ARL-TR-7286 ● May 2015

US Army Research Laboratory

Advanced Video Activity Analytics (AVAA): Human Factors Evaluation

by Patricia L McDermott, Beth M Plott, Anthony J Ries, Jonathan Touryan, Michael Barnes, and Kristin Schweitzer

NOTICES

Disclaimers

ARL-TR-7286 ● MAY 2015

US Army Research Laboratory

Advanced Video Activity Analytics (AVAA): Human Factors Evaluation

Patricia L McDermott and Beth M Plott
Alion Science and Technology

Anthony J Ries, Jonathan Touryan, Michael Barnes, and Kristin Schweitzer
Human Research and Engineering Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
May 2015	Final	September 2013–October 2014

4. TITLE AND SUBTITLE
Advanced Video Activity Analytics (AVAA): Human Factors Evaluation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Patricia L McDermott, Beth M Plott, Anthony J Ries, Jonathan Touryan, Michael Barnes, and Kristin Schweitzer

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
ATTN: RDRL-HRM-A
Aberdeen Proving Ground, MD 21005-5425

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-7286

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
A Human Systems Integration evaluation of the Advanced Video Activity Analytics (AVAA) system was conducted to capture baseline performance and workload with the AVAA system and compare it to performance with advanced AVAA features. This first-year assessment focused on the impact of V-NIIRS (Video National Imagery Interpretability Rating Scale), a widely used scale to evaluate video imagery quality. Experienced analysts searched for targets in full-motion video using AVAA software, both with and without V-NIIRS filter capabilities. Measures of performance included percent of primary targets found, time to find primary target, total targets found, and buttons clicked. Traditional subjective assessments of workload were augmented with continuous physiological and behavioral measurements in order to capture more accurate cognitive state fluctuations during human-system interaction. The findings suggest that analysts were able to identify more targets with the V-NIIRS filter than in the baseline condition in time-pressured situations. The study also developed and implemented a multiaspect approach to estimate operator functional state during system evaluation.

15. SUBJECT TERMS
video analytics, EEG, usability, auditory-evoked potentials, full motion video, Human Systems Integration, AVAA, cognitive workload, physiological measures, human factors, assessment, usability

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Michael Barnes
					19b. TELEPHONE NUMBER (Include area code)
Unclassified	Unclassified	Unclassified	UU	68	520-538-4702

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

Modern warfare is in many ways information warfare. Military success will be determined by the ability to locate, assess, and take action against adversarial forces or terrorists' cells before they can act. The ability to transform information into intelligence is a requisite of information warfare. The analyst must combine his/her understanding with the stream of available information to produce actionable intelligence. With the plethora of information systems available for dissemination at all echelons, too much information is often the problem, not the solution. The Army's transfer to cloud computing both improves the situation and makes information availability more problematic. Cloud computing is more effective and efficient than the current distributed Army networks, and it also makes global information sources and higher-end information processing resources accessible at lower echelons (Keller 2012).

Currently, analysts must manually scan through full-motion videos (FMVs) to find a particular target or activity. They can search for video by geolocation or by time but must watch all of the video to find any features of interest. As a result of the massive amounts of time required to watch all FMVs that are recorded in an area or at a particular time, most video is left untouched and many targets of interest are assumed missed. There is an increasing demand for access to, analysis of, and exploitation of FMV. With so much FMV being recorded and live missions being conducted, forensic analysis suffers because there are too few analysts to perform manual processing, exploitation, and dissemination.

1.2 Advanced Video Activity Analytics (AVAA) Overview

The AVAA system is slated to serve as the sole FMV exploitation capability for the Distributed Common Ground Station-Army. AVAA's objective is to dramatically reduce the analyst's cognitive workload and to enable faster and more accurate production of intelligence products (Swett 2013). The completed version of AVAA will unlock the content of video for high levels of correlation with data across the warfighter enterprise by automatically analyzing, annotating, and organizing massive volumes of video.

AVAA is designed to help analysts collect, analyze, store, and manage FMV data (Fig. 1). AVAA collects FMVs for real-time analysis and forensic investigation. AVAA is used to analyze information by improving the ability to filter, access, and annotate FMVs. AVAA is designed to store and manage the information products so users can quickly find the information for which they are looking. The screenshot in Fig. 2 shows an FMV with a clickable timeline below the video feed and a list of annotations to the right of the screen. AVAA is being developed to work with selected computer vision algorithms (CVAs) that are being developed independently. The CVAs include precision geolocation; detection and characterization of persons, vehicles, and objects; tracking; face detection and recognition; motion stabilization; license plate detection; and metadata resolution.

Collect	Analyze	Store	Manage
• Live data feeds • Forensic data collections	• Detect entities of interest • Create annotations	• Video indexed by metadata and annotations • Cloud capabilities with local control	• Search by annotations • Workflow management tools

Fig. 1 AVAA functionality

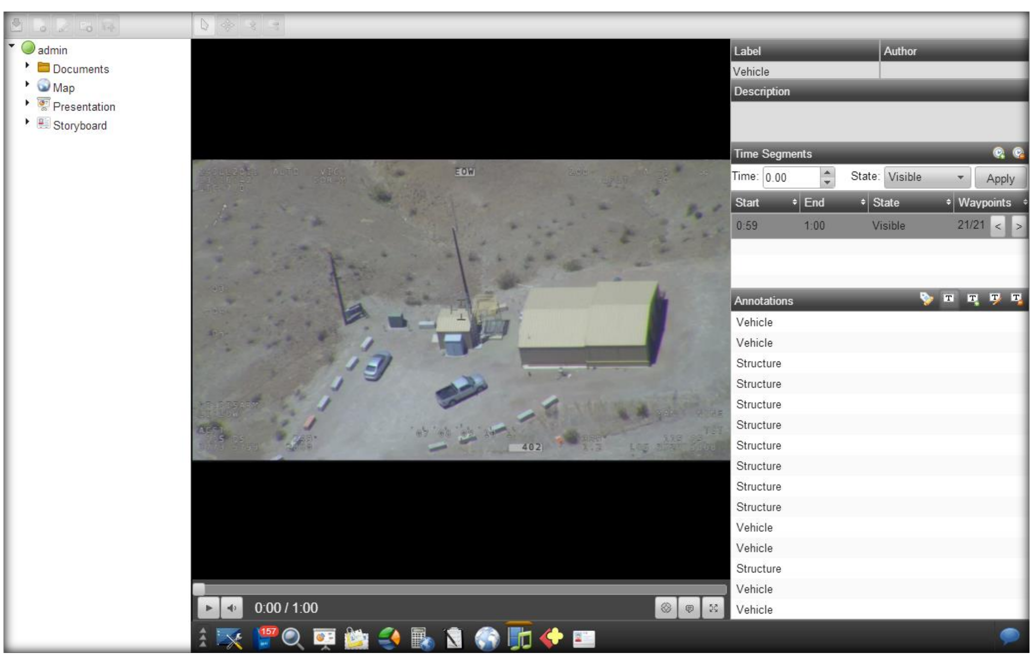

Fig. 2 A screenshot from an early version of AVAA

AVAA will include filtering capabilities to help narrow down the total number of FMVs to be screened and focus on the FMVs that are most likely to contain scenes of interest. One such filter capability is the V-NIIRS (Video National Imagery Interpretability Rating Scale) filter. V-NIIRS is a widely used scale to rate the interpretability of a given image. The V-NIIRS ratings are automatically generated by AVAA. The ratings and examples of targets that can be identified with each rating are shown in Table 1 (Federation of American Scientists 2014). Each frame in the video is given a rating; therefore, a single FMV will have a range of V-NIIRS ratings. The filter returns FMVs that have the requested V-NIIRS rating in at least one frame within the video. In addition to filtering out low-quality videos, the V-NIIRS feature displays a visualization of the changing V-NIIRS rating over the course of an FMV. Fig. 3 shows the V-NIIRS rating graph below the video feed. The graph aligns with the timeline, and analysts can click on a point in the graph to view video of a specific rating. This could be useful in directing analysts to video sections with a higher zoom or focus, which may be due to an object of interest in the field of view.

Table 1 Video National Imagery Interpretability Rating Scale (V-NIIRS)

V-NIIRS Rating	Identifiable Targets
0	• Interpretability of the imagery is precluded by obscuration, degradation, or very poor resolution
1 [over 9.0 m GRD]	• Detect a medium-sized port facility and/or distinguish between taxi-ways and runways at a large airfield.
2 [4.5–9.0 m GRD]	• Detect large static radars • Detect large buildings (e.g., hospitals, factories).
3 [2.5–4.5 m GRD]	• Detect the presence / absence of support vehicles at a mobile missile base. • Detect trains or strings of standard rolling stock on railroad tracks (not individual cars)
4 [1.2–2.5 m GRD]	• Detect the presence of large individual radar antennas • Identify individual tracks, rail pairs, control towers.
5 [0.75–1.2 m GRD]	• Identify radar as vehicle-mounted or trailer-mounted. • Distinguish between SS-25 mobile missile TEL and Missile Support Vans (MSVS) in a known support base, when not covered by camouflage.
6 [0.40–0.75 m GRD]	• Distinguish between models of small/medium helicopters • Identify the spare tire on a medium-sized truck.
7 [0.20–0.40 m GRD]	• Identify ports, ladders, vents on electronics vans. • Detect the mount for antitank guided missiles (e.g., SAGGER on BMP-1).
8 [0.10–0.20 m GRD]	• Identify a hand-held SAM (e.g., SA-7/14, REDEYE, STINGER). • Identify windshield wipers on a vehicle.
9 [less than 0.10 m GRD]	• Identify vehicle registration numbers (VRN) on trucks. • Identify screws and bolts on missile components.

Note: GRD = ground-resolved distance.

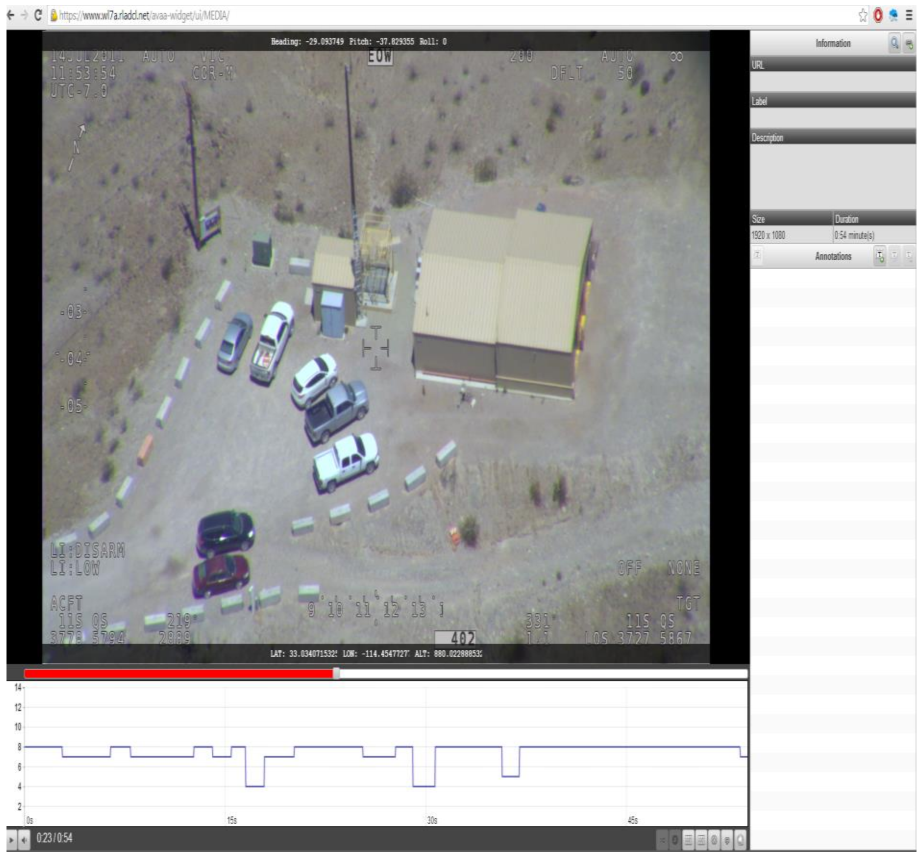

Fig. 3 AVAA screenshot with V-NIIRS rating graph

1.3 Analyst's Task

The imagery analyst job encompasses a wide range of tasks and goals. A representative sample task, the one that was used in the experiment, involves pre-entry phase planning for a secure and stabilization mission in a previously unoccupied country. Entrance into the country will occur in 2 months. Imagery analysts are briefed on the enemy situation, including past and predicted enemy activities, enemy grievances, enemy attack size and operating procedures, weapons, vehicles, and communications. Within the last few months there were numerous general reconnaissance unmanned aerial vehicle (UAV) flights over the area of interest that have not yet been exploited. The brigade commander wants to learn as much as possible about activity and infrastructure in the region before starting detailed planning for the operation. The commander issued a list of essential elements of information (EEI) intended to quickly and effectively expand the unit's knowledge base. The EEI includes infrastructure of military significance (e.g., buildings, compounds, communications facilities, training sites, specialized facilities/sites, motor pools/harbors/docking facilities, secure sites/securing fencing) and activities of military significance (e.g., single vehicles and convoys, tracked vehicles, watercraft, personnel, individuals, and formations, security patrols, and maintenance repairs or support). The brigade commander directed the available imagery be given an initial rapid screening, and observations pertinent to the EEI be annotated with emphasis on capturing location, date and time, and descriptive notes where appropriate. The goal is to screen many videos and capture and annotate observations of potential significance to the brigade mission.

To meet these goals, an analyst searches for video that meets the mission criteria. A list of FMVs that meet the criteria is returned from the search. The analyst selects a video from the list to view. While viewing the video, the analyst uses traditional controls of play, pause, and stop. Fast forward and rewind buttons are currently not available, but analysts can click on any spot in the timeline and the video will jump to that spot. Analysts can click on the timeline to move the video forward in small increments, such as 10 s. Doing this repeatedly is referred to as "scrubbing" forward so that the analyst sees screenshots from the video in quick secession. If the analyst sees something of interest, the analyst annotates it by drawing a rectangle on the entity of interest and typing a label. Once the analyst finishes with the video, he or she can choose another from the list and repeat the process.

1.4 Performance Assessment

The intended impact on the analyst is reduced workload, reduced time to analyze video (and thus increase the amount of video one analyst can exploit), and improved ability to locate targets accurately within the videos. To assess workload, evaluators have traditionally relied on self-assessment questionnaires to provide estimates of cognitive state; however, many self-assessment questionnaires require the operator to be interrupted at discrete times throughout the testing session. Not only does the interruption break mental concentration on the task, but self-reports are not sensitive to fluctuations of cognitive state within a task; they instead provide an average subjective estimate over a length of time. A potential solution to this problem involves the continuous physiological and/or behavioral measurement of task performance.

Physiological and/or behavioral measurements, such as electroencephalography (EEG), eye-tracking, and overt performance (e.g., reaction time and accuracy), have shown reliable, objective quantification of cognitive states associated with workload and fatigue (Berka et al. 2007; Dinges et al. 1998; Dinges and Powell 1985; Johnson et al. 2011; Makeig and Inlow 1993; Stikic et al. 2011). In fact, some evidence suggests that both neural and ocular measurements may be more sensitive to cognitive states like workload when compared to subjective self-reports (Ahlstrom and Friedman-Berg 2006; Peck et al. 2013).

While EEG does show general patterns of neural activity related to cognitive workload across individuals, neural features associated with this construct are often idiosyncratic. Neural classification of cognitive workload and other cognitive states is greatly improved by implementing user-specific models rather than relying on a normative generalized model (Johnson et al. 2011; Kerick et al. 2011; Wilson and Russell 2007, though see Wang et al. 2012 for an exception). The continuous model approach often necessitates the administration of baseline tasks prior to testing in order to create user-state models specific to the operator.

In addition to EEG, eye-tracking measurements provide further objective indices of user state. For example, research has shown that as task demands rise and cognitive workload increases, blink rate and blink duration decrease and fixation frequency (number of fixations/time) increases (Ahlstrom and Friedman-Berg 2006; Van Orden et al. 2001; Wilson 2002). Others have observed changes in pupil diameter as a function of workload, noting decreases in pupil diameter as workload increases (e.g., Backs and Walrath 1992; Van Orden et al. 2001). Using a sustained visual tracking task, Van Orden et al. (2000) found that fixation dwell time and blink duration were highly predictive of task performance such that fixation dwell time decreased and blink duration increased as a function of fatigue-related performance error (Van Orden et al. 2000). In line with EEG findings, individualized models of eye activity tend to be better predictors of performance relative to a general model (Van Orden et al. 2000). Together, these findings indicate that multiple eye-tracking metrics are valuable in assessing the cognitive state of an operator.

This project presents a proof-of-concept approach to assessing operator functional state as a means to evaluate system design. We focused on cognitive workload during FMV analysis. Operators performed a target search task while evaluating FMV using 2 different software implementations. We evaluated both continuous and discrete electrophysiological estimates of cognitive workload. Additionally, we collected ocular metrics and behavioral responses to a secondary task.

1.5 Project Goals

This report describes a human factors evaluation of AVAA to empirically validate the filtering capabilities of AVAA for performance improvement and for workload reduction. The human factors assessments are ongoing evaluations of different stages of AVAA both to improve the operator’s interaction with the system and to continually enhance and evaluate AVAA as it is being developed.

The human factors study included empirical evaluation and user feedback. In the empirical evaluation, researchers captured user actions, physiological measures, and system usability during realistic scenario-based operations. Two data collection events took place to obtain baseline data and preliminary data on the V-NIIRS filter, a widely used scale to evaluate video imagery quality. A pilot test in April 2014 set the stage for a more formal assessment in June. The purpose of both the pilot and the formal assessment was to better understand the operator's workload and performance and to capture design recommendations in terms of capabilities, interface improvements, and any problems encountered in the assessment process.

2. Pilot Experiment

The pilot test was conducted at the Experimentation and Analysis Element (EAE) at Ft. Huachuca from 14 to 17 April 2014. Data collection was a joint effort between the US Army Research Laboratory, Alion Science and Technology, and AVAA contractors from Chenega and EOIR corporations.

2.1 Objective

Our objective in the pilot was to try out the data collection software, experimental design, EEG, and survey forms and to collect design recommendations from active duty imagery analysts stationed at the US Army Intelligence Center of Excellence (ICoE) at Ft. Huachuca.

2.2 Method

2.2.1 Experimental Design

The experiment was a 2×2 mixed design. Quality Filter was a within-subjects variable with 2 levels: 1) a Baseline condition in which V-NIIRS was not used and 2) a V-NIIRS condition. The V-NIIRS condition provided an additional filter to narrow down possible FMVs by video quality as well as a clickable graph of V-NIIRS ratings that was visible when viewing the FMVs. The Presentation Order was a between-subject variable. All participants experienced both conditions; however, half the subjects saw scenario A under the V-NIIRS condition and then saw scenario B under the Baseline condition. The other half of the subjects saw the reverse pairing (scenario B with V-NIIRS; A with Baseline). The conditions were counterbalanced to control for the order in which the scenarios were presented to participants.

2.2.2 Participants

There were a total of 6 participants: 2 35G (enlisted) trained analysts, 3 warrant officer analysts, and 1 civilian not trained in imagery analysis. The civilian is included as a pilot participant because the civilian was one of the 2 EEG participants. An additional 35G noncommissioned officer (NCO) familiar with AVAA gave verbal feedback. The analysts had between 1.3 and 7 years of experience in the Imagery Analysis military occupational specialty (MOS) (M = 4.67 years, SD = 2.17). Every analyst had operational imagery analysis experience.

2.2.3 Equipment and Materials

2.2.3.1 AVAA Workstations

The data collection took place at the US Army ICoE EAE at Ft. Huachuca, AZ. The laboratory consisted of 5 laptop workstations each with a full-size stand-alone 20-inch monitor, keyboard, and mouse. The video consisted of data supplied by Yuma Proving Ground, the Unmanned Aerial System program office at Redstone Arsenal, and other data sources identified by the EOIR Corporation. Each video had a time/date stamp, geolocation information, and a V-NIIRS number for the target of interest.

2.2.3.2 EEG and Eye Gaze Data Collection Suite

EEG data were acquired (sampling rate 256 Hz) from the B-Alert x24 Wireless Sensor Headset using the B-Alert software package (Advanced Brain Monitoring, Carlsbad, CA) (Fig. 4). Wireless EEG signals were sent via Bluetooth to an external synching unit, which connected to a data acquisition laptop through USB. In addition to the scalp electrodes, 2 external input channels were used to acquire electrocardiogram data.

Eye movement data were recorded using the Tobii X120 eye-tracker. Data from each eye were sampled at 120 Hz and acquired using custom software with the Tobii Software Development Kit. Data were recorded on the same machine as the EEG through a custom Ethernet connection. Prior to testing, each operator performed a 9-point calibration. Eye tracking data were used to measure fixation and blink frequency as well as provide estimates of gaze distribution. Participants were asked to rate their subjective cognitive state (e.g., workload) at the conclusion of each scenario.

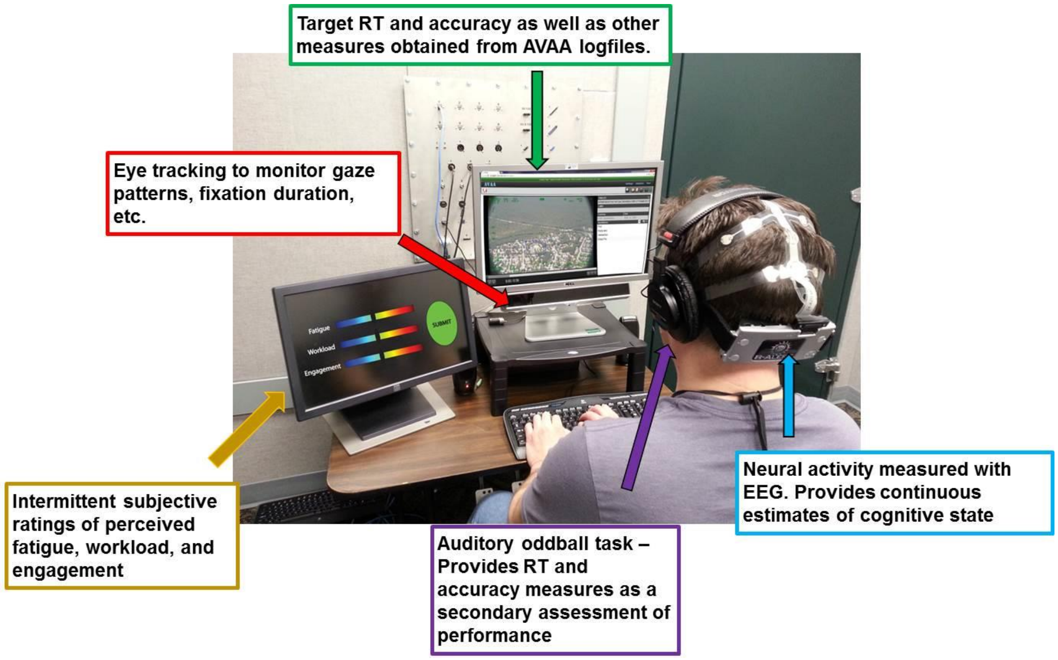

Fig. 4 EEG data collection station

2.2.3.3 Forms and Questionnaires

Four questionnaires were used:

- A demographics form queried age, gender, formal education level, MOSs (present and past), time in those MOSs, time actually performing the relevant MOS duties, whether eyeglasses were needed, and other experience relevant to AVAA operations.
- The Short Stress State Questionnaire (SSSQ) captured each analyst's self-assessment of interest in the task, level of focus, and tiredness for that particular day.
- NASA TLX Part 1 captured subjective ratings of mental demand, physical demand, temporal demand, performance, effort, and frustration. Part 2 was used to assess the relative importance of the 6 factors on the experienced workload.
- A Usability Questionnaire captured analysts' ratings of AVAA software clarity and learnability, actions and memory load required, user guidance, and training. Ratings were labeled "strongly agree," "agree," "disagree," "strongly disagree," and "not applicable."

See Appendix A for all 4 questionnaires.

2.2.4 Procedure

2.2.4.1 Non-EEG Participants

Participants completed a consent form and demographic form. AVAA personnel conducted a short group training session to familiarize participants with the AVAA software functionality. Participants then used AVAA during realistic, scenario-based missions to search, select, view, and annotate FMV. Participants did one scenario set in the Baseline condition and one scenario set in the V-NIIRS condition. A scenario set included 5 tasks, each with a different time, date, V-NIIRS range (if applicable), and target to locate.

In the baseline condition, the participants searched through videos in specific time frames (e.g., 0600 to 0800 h on 17 November 2013). For the filtered conditions, the V-NIIRS filter was used in the search criteria to filter out low-quality imagery for the time period chosen. Participants were told to search for a specific target within each task and to use the annotation tools to describe the target. There was no time limit for the tasks. After completing the scenario set in their first condition, participants completed a paper-based version of the NASA TLX: Part 1. After completing the second condition, participants completed Parts 1 and 2 of the NASA TLX. Although there was disparity in the times among participants, the participants took approximately an hour to finish the exercise.

2.2.4.2 EEG Participants

Two participants were fitted with EEG equipment and performed preliminary tasks prior to learning and using the AVAA software. The number of EEG participants was limited because only one EEG station was available. Additional EEG stations would have facilitated running additional EEG participants. While wearing the EEG system, participants performed a psychomotor vigilance task (PVT) and 2 resting tasks, one with eyes open and one with eyes closed. During the PVT, participants made a forced-choice response (2 alternatives) to a colored shape appearing on the computer monitor. During the eyes open and eyes closed tasks, participants made a speeded detection response to a single luminance change on the monitor (eyes open) or an auditory tone (eyes closed). EEG was recorded during these baseline tasks to create an individualized model for each subject. These models serve as the basis for cognitive state estimation during the experiment. Participants also performed an eye-tracking calibration procedure requiring them to fixate on a series of dots within a pattern presented on the computer monitor. The extra EEG tasks and model building phase took approximately 1 h.

EEG participants then attended the group training and completed identical AVAA scenarios as the non-EEG participants. EEG participants performed a simple auditory target discrimination task (the auditory "oddball" task) concurrently with the target identification task. The auditory oddball task required participants to make a speeded response by pressing a button on a touch screen monitor in response to a specific auditory stimulus (the "oddball" tone) that occurred in the midst of distractor auditory stimuli. This type of task has proven effective in discriminating levels of cognitive workload (Allison and Polich 2008; Miller et al. 2011).

2.2.5 Metrics

Performance metrics for each scenario included the number of FMVs returned (i.e., the number of videos that met the search criteria), the number of FMVs viewed, whether the primary target was found, the time it took to find the primary target, and the number of interface buttons clicked while conducting the task. With the exception of the button clicks, all performance metrics were manually collected by experimenters. Button clicks were automatically logged for all 6 participants. For 2 of the participants (P1 and P5), EEG and eye-tracking data were collected. Usability surveys, the NASA-AMES TLX workload scale, demographics, and debriefing data were collected for the 5 analyst participants.

2.3 Pilot Results

2.3.1 Performance Metrics

2.3.1.1 Impact of Filter on Workflow

The baseline condition had a mean of 12.13 FMVs returned from their search. The V-NIIRS condition had a mean of 9.30 videos—a reduction of 23%. In the baseline condition, participants viewed a mean of 5.19 videos. In contrast, participants in the V-NIIRS condition viewed a mean of 2.90 videos—a reduction of 44%.

2.3.1.2 Impact of Filter on Performance

The 2 primary metrics centered on task time and accuracy. This included percentage of primary targets found and time to find the primary target. The descriptive statistics show that in the V-NIIRS condition, participants were more successful and faster at finding targets (Table 2). In the V-NIIRS condition, participants found a mean of 86.96% of primary targets—an increase of 11% more primary targets found than in the baseline condition. Participants were 11%

faster in finding and annotating targets in the V-NIIRS condition. While false positives were possible if an analyst incorrectly identified an entity, no false positives were observed. Note that the standard deviations for each metric are high, indicating that the differences are not likely to be statistically significant.

Table 2 Task time and accuracy

	Primary Time (min)		Primary Found (%)	
	Mean	**St. Dev.**	**Mean**	**St. Dev.**
Baseline	7.08	4.26	78	42
V-NIIRS	6.30	4.60	87	34

2.3.1.3 Button Clicks

The button clicks were analyzed to characterize the way in which participants used the system. Most of the button clicks could be classified into 2 categories: 1) playing and advancing the video and 2) creating and saving annotations (Fig. 5). The search button clicks were not recorded in the data log for the April test. Playing and advancing the video included play, pause, scrub forward, and scrub backwards. There was a negligible number of other clicks that did not fit into these categories (e.g., mute) that were not analyzed. The number of annotation clicks ranged from 11 to 32 with a mean of 20 clicks (SD = 8.75). The number of play/advance clicks had the most variability, ranging from 304 to 4,813 clicks with a mean of 2,149 clicks (SD = 1,569.7). The analysts each had over 1,000 clicks during the 10 scenarios, while the civilian had only 316 total clicks. This provides evidence that trained analysts approached the task differently and clicked much more frequently to accomplish the tasks. On average, the play/advance clicks made up 99% of the total clicks. Keyboard alternatives for clicking were not observed for play and annotation actions.

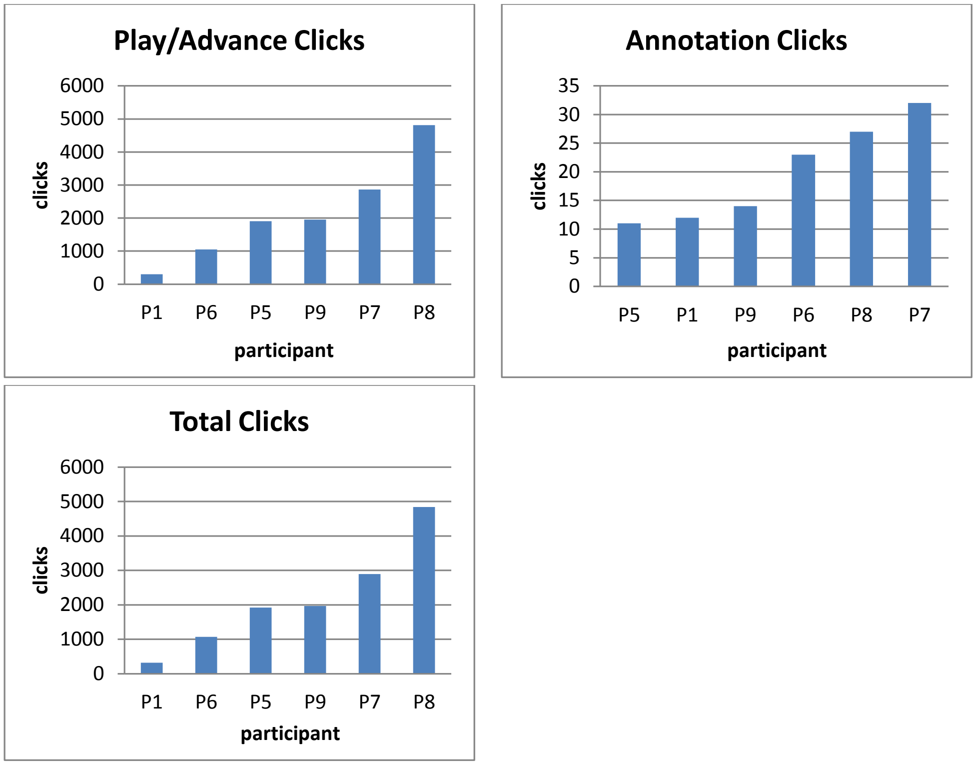

Fig. 5 **Clicks by participant for categories of annotate, play/advance, and total**

2.3.2 Questionnaires

2.3.2.1 NASA TLX

The NASA TLX is a subjective workload scale that is widely used by researchers (Hart and Staveland 1988). The raw responses vary between 1 and 20 and are then weighted by individual. The weighted workload ratings for the Baseline and V-NIIRS conditions are shown in heat maps in Tables 3 and 4, respectively. The warmer the color is, the higher the workload rating. Note that for the Performance scale, higher ratings are desirable, as they indicate that analysts were highly satisfied with their performance. High ratings can be seen in Mental Demand (MD), Performance (P), and Frustration (F). As expected, Physical Demand (PD) had consistently low workload ratings. The overall weighted workload rating was 8.77 (SD = 3.76) for the Baseline condition and 10.10 (SD = 3.28) for the V-NIIRS condition. In comparing the 2 heat maps, the V-NIIRS condition appears to have lower temporal demand, higher performance ratings, and lower effort. The weighted workload for each category by condition is shown in Fig. 6.

Table 3 Heat map of workload ratings for Baseline

Participant	Instance	Condition	MD	PD	TD	P	E	F
5	2	Baseline	42	6	19	20	17	75
6	1	Baseline	12	0	12	42	4	18
8	1	Baseline	0	0	8	20	12	15
9	1	Baseline	40	0	60	12	7	44
11	2	Baseline	18	4	14	85	10	42

MD = Mental Demand; PD = Physical Demand; TD = Temporal Demand; P = Performance; E = Effort; F = Frustration

Table 4 Heat map of workload ratings for V-NIIRS

Participant	Instance	Condition	MD	PD	TD	P	E	F
5	1	V-NIIRS	48	3	13	52	11	85
6	2	V-NIIRS	8	0	4	60	2	3
8	2	V-NIIRS	31.5	0	9	52	6	60
9	2	V-NIIRS	28	0	44	22	7	40
11	1	V-NIIRS	33	4	8	80	5	39

MD = Mental Demand; PD = Physical Demand; TD = Temporal Demand; P = Performance; E = Effort; F = Frustration

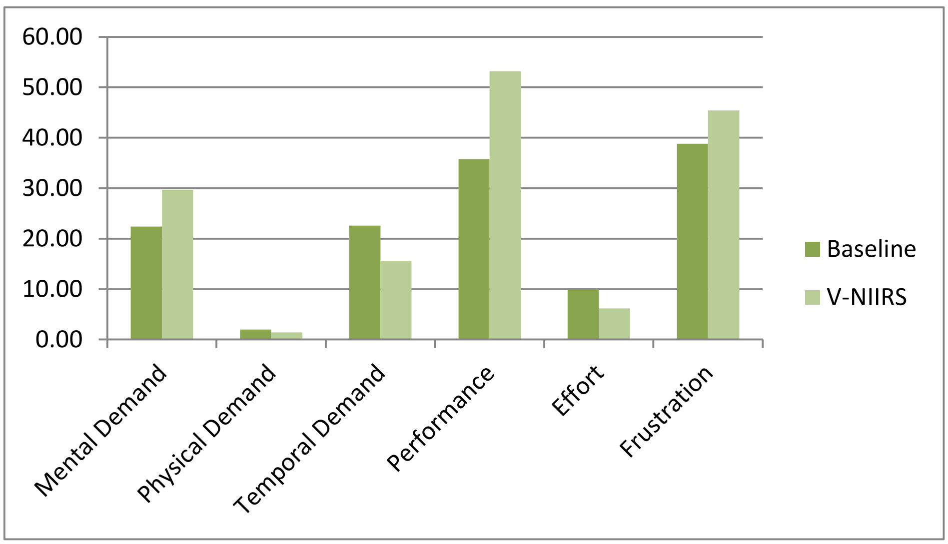

Fig. 6 Weighted NASA-TLX workload ratings by condition for the pilot experiment

2.3.2.2 Short Stress State Questionnaire (SSSQ)

The SSSQ consists of 24 items that are rated on a scale from 4 (Extremely) to 0 (Not at all). Ratings are summed to create scores for 3 different subscales: Engagement, Distress, and Worry. Each subscale has 8 associated questions for a maximum possible score of 32. The mean Engagement score was 26.40 (SD = 7.44) with a range from 14 to 32. The mean Distress score was 7.40 (SD = 5.68) with a range from 0 to 13 and the mean Worry score was 7.00 (SD = 5.66) with a range from 2 to 16. Figure 7 shows the subscale scores by participant.

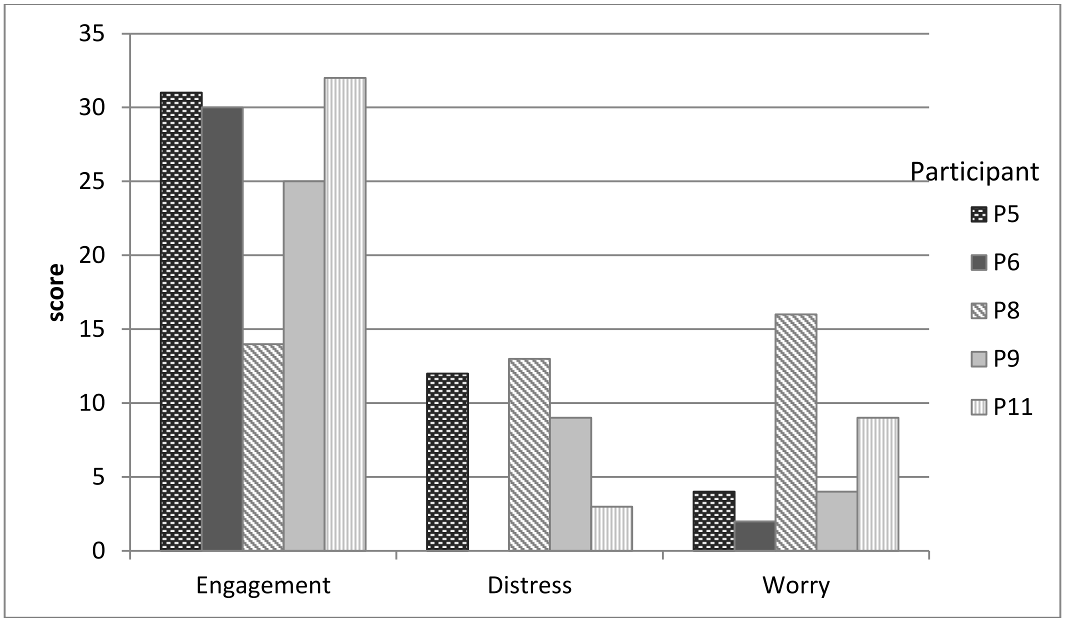

Fig. 7 Short Stress State Questionnaire (SSSQ) ratings of engagement, stress, and worry by participant for the pilot experiment

2.3.2.3 Usability

There was a usability survey of 44 items asking questions about the quality of the interface and the capabilities of the systems to conduct 35G missions. The scale ranged from 5 (strongly agree) to 1 (strongly disagree). The questions were designed so that "agree" indicated a good/beneficial feature. The average score over 43 items was 3.56 (SD = 1.09), with 43% of the ratings favorable (a rating of a 4 or 5). The questions were categorized into 3 groups: Interface, Functionality, and MANPRINT. The 17 questions in the Interface category had a mean of 3.51 (SD = 0.37). Ratings that stood out with disagreements or agreements are shown in Table 5.

Table 5 Interface statements

Disagree Ratings	Agree Ratings	Statement
2	. . .	Data shown on the display screen are always in the format I need.
3	. . .	It is easy for me to tell what data or files I am actually transmitting.
. . .	5	It is relatively easy to move from one part of a task to another.
. . .	5	It is easy to acknowledge system alarms, signals, and messages.

The 17 questions in the Functionality category had a mean of 3.47 (SD = 0.54). Ratings that stood out with disagreements or agreements are shown in Table 6. The following 6 statements were rated as not applicable by at least 3 out of 5 analysts:

- AVAA does not interfere with other programs I use.
- Importing data into the system is easy.
- Exporting data out of the system is easy.
- I can easily get a printed copy of the screen when I need it.
- I rarely have to reenter data that I know is already available to AVAA in other files.
- If AVAA rejects my input, it always gives me a useful feedback message (i.e., tells me why and what corrective action to take).

Most of the statements rated not applicable were not exercised during the scenarios.

Table 6 Functionality statements

Disagree Ratings	Agree Ratings	Statement
2	. . .	AVAA provides all the information I need to do my work.
. . .	4	When a keystroke (or mouse click) does not immediately produce the response I expect, the software gives me a message, symbol, or sign to acknowledge my input.

The 9 questions in the MANPRINT category had the highest mean of 4.00 (SD = 0.84). Ratings that stood out with disagreements or agreements are shown in Table 7. The statement "Compared to my current method of exploiting imagery, AVAA does not affect my workload" could be interpreted as AVAA either increasing or decreasing their workload.

Table 7 MANPRINT statements

Disagree Ratings	Agree Ratings	Statement
2	. . .	Compared to my current method of exploiting imagery, AVAA does not affect my workload.
2	. . .	I have encountered no design or ergonomic issues with regard to system hardware.
. . .	4	The number of personnel available in my unit/section is adequate to support full AVAA operations.
. . .	5	I have the appropriate MOS to complete all assigned tasks.
. . .	5	There are no physical limitations (color vision, hearing, etc.) that prevent me from completing tasks.
. . .	5	The walk-through training gave me sufficient guidance so that I was able to complete my assigned task.
. . .	4	Learning to use this software is easy.
. . .	5	I feel confident in my ability to complete my assigned task using AVAA.

The 44th rating queried how long it would take before the analyst would be comfortable in the use of AVAA to complete job tasks. The options were less than 1 month, 2–3 months, 4–6 months, 7–12 months, and more than 12 months. Three analysts felt that it would take less than a month to become comfortable with using AVAA in order to conduct their missions. One analyst felt it would take 2–3 months and another felt it would take 7–12 months.

2.3.3 Observations

The debriefing proved very useful for possible design improvements. In general, the participants felt that AVAA was a useful tool for FMV intelligence analysis and are looking forward to the advanced versions. The comments and observations were categorized as bugs (4), capability requests (32), and process feedback (11) and listed in Appendix B. Several capability requests dealt with the list of returned videos and being able to differentiate the videos from each other and determine if a video had been viewed. Some capability comments dealt with the viewing of videos. Analysts felt it was important to be able to view the videos in faster than real time. Manually "scrubbing" the video moved to a farther point in the video. This allowed the user to move through the video more quickly, but it meant that there were parts of the video that were scrubbed past and never viewed. These video snippets could have contained useful information. Some analysts dealt with this by scrubbing very slowly (e.g., only moving forward a small amount of time), but this caused the user to click on the interface many times—over 1,000 clicks in the course of the scenarios. This can lead to fatigue, frustration, and missed targets. Thus, being able to move through the video rapidly but also being able to revisit sections of interest would increase the versatility of the interface.

Some comments dealt with features that already exist but were not exercised during the scenarios, such as the ability to resize the window components. Others dealt with planned enhancements to AVAA, such as the ability to export still pictures from the FMVs. The users offered suggestions on how to make navigation and data entry more user friendly. One suggestion was to have the filter/search criteria visible while a video was playing. Users requested more annotation tools, including the ability to use different colors and different shapes, and the ability to drag and drop MIL-STD-2525 (2008) symbols onto video. The analysts also wanted more map features, such as map layers and the ability to see which part of the terrain was currently on video.

The list of capability requirements is being used by the software engineers to make interface changes and as a to-do list for changes in future software builds.

As a result, future priorities will be directed toward more user-friendly interfaces based on the analyst's feedback.

The process requests were lessons learned by the researchers on how the experimental design, scenarios, and procedures could be improved. These were addressed in the formal June evaluation. For example, an operational context describing the mission and the Commander's Priority Information requirements was provided to analysts.

2.4 Pilot Discussion

There were not enough subjects to warrant an analysis of variance; therefore, all comparisons are based on descriptive statistics and do not indicate statistical significance. The pilot data showed that V-NIIRS reduced the work to be done. Analysts in the V-NIIRS condition were given 23% fewer videos to exploit, and on average they viewed 44% fewer videos. Participants in the V-NIIRS condition found 11% more targets even though they viewed fewer videos.

One usability issue was the large number of button clicks by the participants. This was largely due to participants scrubbing the video to advance to a later frame. AVAA is not supporting the task of quickly exploiting video in faster than real time. We recommend allowing analysts to view the video in faster than real time (e.g., 2×, 4×, 8×, 16×, 32× speed). This will allow them to see all the video as opposed to skipping frames, and it will reduce the number of interface clicks. In general, the analysts were engaged in the task. Higher workload was observed on the mental demand, performance, and frustration scales of the self-assessment tool. Workload was 15% higher in the V-NIIRS condition than the baseline condition, perhaps indicating greater engagement in the V-NIIRS conditions.

In terms of experimental design, we concluded that a number of changes were necessary for the June data collection. The analyst's task was expanded, requiring him or her to search for a list of targets of opportunity (essential intelligence elements) as well the single target of interest for each scenario. This was more realistic in terms of actual mission procedures and it made data collection more efficient. The operational context of the scenarios was made more specific, and each of the scenarios was limited to 10 min to ensure that the baseline and V-NIIRS conditions were run under the same constraints.

3. June Data Collection Event

3.1 Objective

The June event was conducted 10–12 June 2014 in the EAE facility at Ft. Huachuca. The objective was to empirically validate the filtering capabilities of AVAA for performance improvement and for workload reduction. Many of the techniques used in the pilot study were replicated in the June assessment. The following descriptions only mention differences between the 2 data collections. The primary differences were the greater level of control in the June event and the more specific operational context given to the analyst during instructions and mission tasking.

3.2 Method

3.2.1 Experimental Design

The experimental design, equipment, materials, and metrics were identical to the pilot study. There were 2 types of scenarios. Intelligent preparation of the battlefield scenarios had stationary primary targets. Moving target scenarios had mobile primary targets, such as vehicles and watercraft. The V-NIIRS-filtered FMVs were viewed to ensure that the V-NIIRS filter did not screen out primary targets. The order of conditions and scenario was counterbalanced and is shown in Table 8.

Table 8 Presentation order for conditions and scenarios

Participant	First Condition	Scenarios	Second Condition	Scenarios
1	V-NIIRS	IPB	Baseline	MT
2	Baseline	MT	V-NIIRS	IPB
3	Baseline	IPB	V-NIIRS	MT
4	V-NIIRS	MT	Baseline	IPB
5	V-NIIRS	IPB	Baseline	MT
6	Baseline	IPB	V-NIIRS	MT
7	Baseline	MT	V-NIIRS	IPB
8	V-NIIRS	MT	Baseline	IPB

IPB = preparation of the battlefield; MT = moving target.

3.2.2 Participants

Because of the constraints at the NCO academy, we were only able to run 6 active duty imagery analysts (350G and 35G) and 2 former analysts for a total of 8 participants. The pool consisted of a chief warrant officer 3 with 12 years of experience, 2 sergeants (E-5), 2 staff sergeants (E-6), and 3 sergeants first class (E-7), all with recent combat experience. The analysts had between 5 and 18 years of experience in the Imagery Analysis MOS (M = 10.2, SD = 4.6).

3.2.3 Procedure

Each analyst filled out a consent form and demographics form. As before, the analyst was trained to use the AVAA software controlling the FMVs to locate targets of interest. There were 3 data collection stations at the EAE, consisting of laptop computers with one being used for EEG data collection. Each station had a data collector to note any unusual occurrences, manually log data, and answer questions during the assessment. All the computers were loaded with AVAA software and videos collected from Yuma Proving Ground, chosen so that each scenario had various elements of military intelligence significance. The analysts were given an operational context to read and were instructed to find a specific target in each scenario; they were also given a list of possible targets that were deemed of intelligence significance and told to report their attributes using the annotation tools. Four of the 8 participants completed the scenarios while using the EEG and eye-tracking equipment. Each analyst was given 4 scenarios to search through and given a short synopsis of the importance of the operational tasking for each scenario. They saw 2 scenarios in the baseline condition and 2 that were filtered using V-NIIRS cut-offs. As in the pilot test, scenario–condition pairings were counterbalanced between subjects. To control for individual differences and differences in the number of videos between conditions, the analyst were given 10 min to complete each scenario, limiting the assessment duration to 40 min. After each session, the analyst was debriefed and filled out a usability survey and a NASA-AMES TLX subjective workload form.

3.2.4 Metrics

Performance metrics for each scenario included the number of FMVs returned (i.e., the number of videos that met the search criteria), the number of FMVs viewed, whether the primary target was found, the time it took to find the primary target, total targets found, and the number of interface buttons clicked while conducting the task. With the exception of the button clicks, all performance metrics were manually collected by experimenters. Data on annotation accuracy and time were not recorded for one participant (P7) so the performance data only reflects 7 participants. Button clicks were automatically logged for all 6 participants. For 4 of the participants (P1, P6, P7, and P8), EEG and eye-tracking data were collected. Usability surveys, the NASA-AMES TLX workload scale, demographics, and debriefing data were collected for all participants.

3.3 Results

3.3.1 Performance Metrics

3.3.1.1 Impact of Filter on Workflow

The baseline condition had a mean of 14.07 FMVs returned from their search. The V-NIIRS condition had a mean of 6.27 videos—a reduction of 55%. In the baseline condition, analysts viewed a mean of 5.36 videos. In contrast, analysts in the V-NIIRS condition viewed a mean of 3.73 videos—a reduction of 30%.

3.3.1.2 Impact of Filter on Performance

The 3 primary metrics were percentage of primary targets found, total targets found, and time to find the primary target. The descriptive statistics show that analysts were more successful but slower at finding targets in the V-NIIRS condition (Table 9). In the V-NIIRS condition, analysts found a mean of 80% of primary targets—an increase of 40% more primary targets found than in the baseline condition. Analysts in the V-NIIRS condition also found and annotated 16% more total targets than in the baseline condition. Because they found and annotated many more targets in the V-NIIRS conditions, the mean time to locate the primary targets they *were* able to find was actually faster in the baseline conditions (2.5 min compared to 6 min for the V-NIIRS).

Table 9 Task time and accuracy

	Primary Time (min)		**Primary Found (%)**		**Annotations (count)**	
	Mean	**St. Dev.**	**Mean**	**St. Dev.**	**Mean**	**St. Dev.**
Baseline	2.55	1.24	57	51	5.57	2.82
V-NIIRS	5.97	2.26	80	41	6.47	3.18

Descriptive statistics were calculated to compare the performance of the 2 analysts with the EEG to the 6 analysts without the EEG. The primary time, primary found, and total annotations of analysts with the EEG were within 7% of those without the EEG, providing evidence that wearing the EEG did not impact performance.

3.3.1.3 Impact of Experience

We examined the correlation between performance and experience using the Pearson product-moment correlation coefficient. There was a positive correlation between MOS experience and primary time, r = 0.84 (Fig. 8). There was a negative correlation between MOS experience and primary found, r = –0.64

(Fig. 9). In both cases, more years of experience in the imagery analysis MOS was associated with poorer performance. There was a negative correlation between MOS experience and videos viewed, showing that more experienced operators tended to view fewer videos, $r = -0.79$ (Fig. 10).

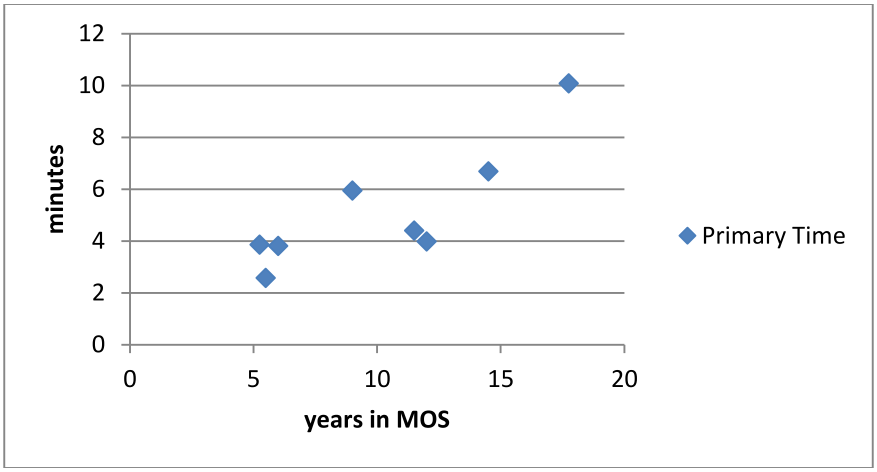

Fig. 8 Time to find primary target by MOS experience

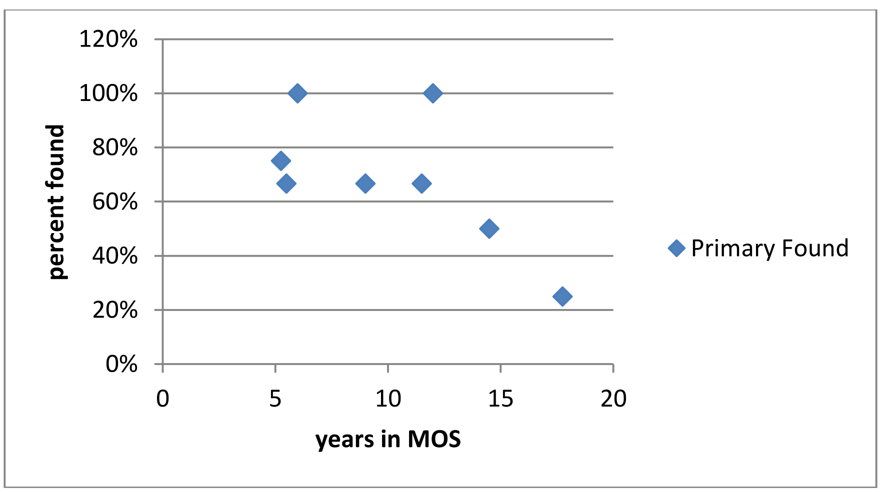

Fig. 9 Primary targets found by MOS experience

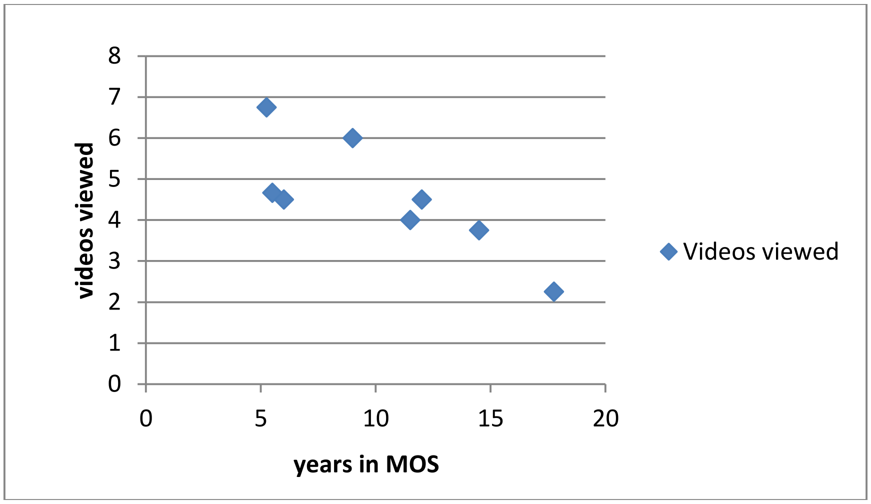

Fig. 10 Videos viewed by MOS experience

3.3.1.4 Button Clicks

The button clicks were analyzed to characterize the way in which analysts used the system. Most of the button clicks could be classified into 3 categories: 1) conducting a search, 2) playing and advancing the video, and 3) creating and saving annotations (Fig. 11). (The logging program was updated after the pilot to also capture search clicks.) Playing and advancing the video included play, pause, scrub forward, and scrub backwards. There was a negligible number of other clicks that did not fit into these categories (e.g., mute) which were not analyzed. The number of search clicks ranged from 12 to 30 with a mean of 17 clicks (SD = 6). The number of annotation clicks ranged from 52 to 149 with a mean of 93 clicks (SD = 32). The number of play/advance clicks had the most variability, ranging from 419 to 10,882 clicks with a mean of 4,404 clicks (SD = 4,342). On average, the play/advance clicks made up 98% of the total clicks.

Five of the 8 analysts had over 7,000 clicks during the 4 scenarios, most of them associated with play/advance. These high clickers made between 7,255 and 10,982 clicks (M = 8933, SD =1,411.64) while the other 3 participants made between 484 and 754 clicks (M = 810, SD = 357.31). On average, the high clickers showed a trend of better performance and more experience. The high clickers found 60% more primary targets, found them 40% faster, and made 78% more annotations than the 3 "low clicker" participants. The high clickers had 3.7 fewer years in the Imagery Analysis MOS (a difference of 30%) and 2.23 fewer years of operational experience (a difference of 33%).

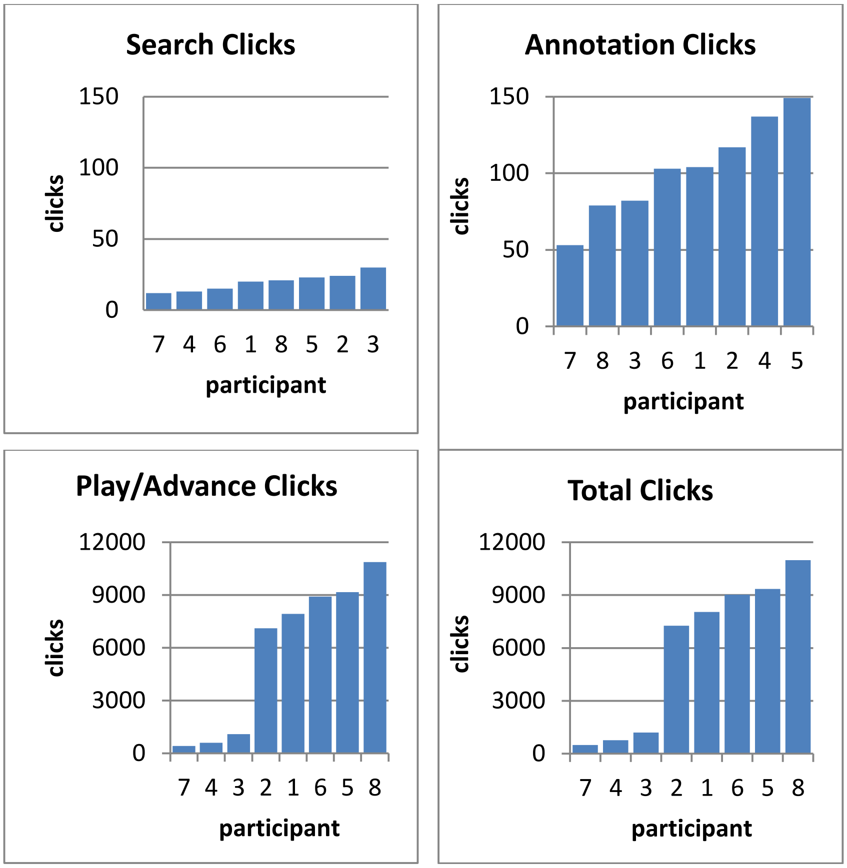

Fig. 11 Clicks by participant for categories of search, annotate, play/advance, and total

We examined the correlation between button clicks and performance. Correlations with search clicks were not examined because extra search clicks may have been due to a software bug in the search process. For annotation clicks, it was not surprising that they were positively correlated with total annotations (r = 0.73). Play/advance clicks were positively correlated with total annotations (r = 0.60) and primary found percentage (r = 0.51). Play/advance clicks were negatively correlated with primary time (r = –0.54).

3.3.2 Behavioral, Neural, and Ocular Metrics for EEG Participants

Only 1 participant could be run using the EEG system in each session. Because of the small number of participants, the data was collapsed and analyzed together. Thus, the data presented in this section is based on 6 participants. Two were from the pilot experiment and 4 were from the formal experiment in June 2014.

3.3.2.1 Electrophysiology

Evoked Potentials

Segments of EEG (1200 ms) were extracted from the continuous EEG signal and time-locked to the onset of the standard and target (oddball) auditory stimuli. These segments, or epochs, were averaged to create event-related potentials (ERPs) for both standard and target/oddball stimuli using a 200-ms baseline (i.e., 200 ms before the onset of the auditory stimulus) and 1,000-ms post-stimulus. In other words, we marked the point in time when a stimulus occurs and averaged a large number of trials so that everything that happened at a relative time (in this case 200-ms pre-stimulus and 1,000-ms post-stimulus) is averaged with everything else that occurred at that relative time. This averaging process filtered out brain activity not related to the appearance of the stimulus. This was done for each target search mission within each condition (Baseline and V-NIIRS). While ERPs were generated for both the target (oddball) and standard auditory stimuli, the target stimuli presented in the auditory task were primarily used as a behavioral performance metric. We focused on the ERPs from the frequent standard stimuli as they provided more samples. We evaluated the amplitude of the N1 ERP component (the first negative-going component) evoked by the auditory stimuli.

The ERPs evoked by the frequent standard stimuli in the secondary task were similar between the Baseline and V-NIIRS conditions (Fig. 12). There was a small trend for the amplitude of the N1 component of the ERP over electrode Cz being slightly larger in the V-NIIRS with respect to the Baseline condition. There was a convincing difference in the N1 component when comparing 2 blocks within the V-NIIRS condition: one in which the operators failed to respond to the auditory targets and another when they were fully engaged in the auditory tasks and successfully responded to all auditory probes. A substantially larger N1 component was found when the operators were actively engaged in auditory tasks compared to when they were strictly focusing on the target search task and ignoring the auditory probe stimuli (Fig. 13). This evidence provides support for the use of auditory probe stimuli to estimate user engagement in a separate task.

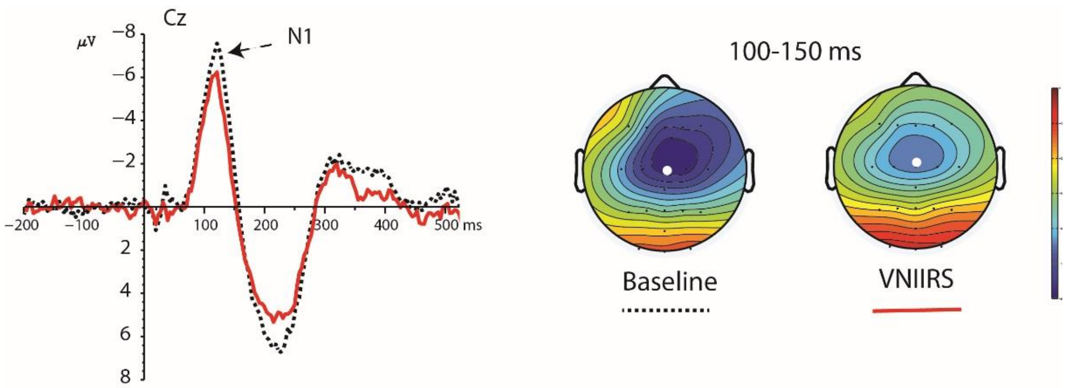

Fig. 12 Auditory-evoked potentials. Left) Auditory N1 component over electrode Cz from standard tones in the Baseline and V-NIIRS conditions. Right) Topographical voltage maps highlighting the scalp distribution of the N1 peak 100–150 ms post-stimulus onset.

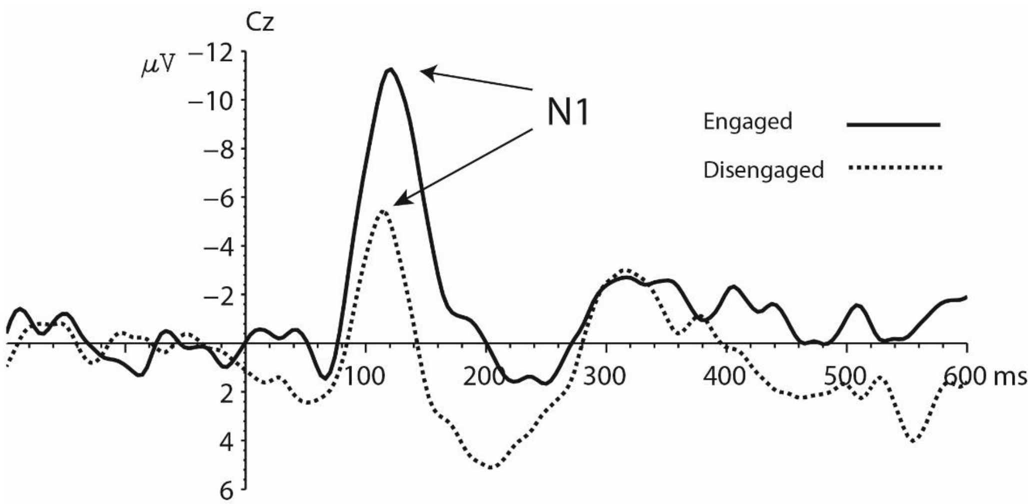

Fig. 13 Auditory-evoked potentials during engaged and disengaged states from operator S05

Workload Classification

Tables 10 and 11 show the average probability of high workload for each mission in the Baseline and V-NIIRS conditions, respectively. These data are derived from the B Alert workload classification model based on the EEG and on average show no difference between the 2 conditions. The top portion of Fig. 14 shows the continuous estimate of workload across all missions, highlighting the fluctuations of workload over time for one analyst. The bottom portion of Fig. 14 shows the cumulative sum of the standardized (Z-scored) workload probability scores over the course of the test. Scores were standardized using the mean and standard deviation from both the Baseline and V-NIIRS conditions. The data depict how workload changed over time with respect to the average of all the missions for that particular analyst. Similar workload estimates between the Baseline and V-

NIIRS conditions were obtained for each operator but have not been analyzed to date. The increased workload in this mission was likely due to a software malfunction at the start of this mission and may reflect neural processes related to a combination operator frustration and workload.

Table 10 Probability of high workload in the Baseline condition for each mission

Participant	M1	M2	M3	M4	M5	Average
S1111	0.60	0.63	0.64	0.62	0.64	0.63
S2222	0.57	0.56	0.58	0.59	0.57	0.58
S0008	0.69	0.73	0.70	…	…	0.71
S0006	0.57	0.57	…	…	…	0.57
S0007	0.68	0.68	…	…	…	0.68
S0001	0.71	0.71	…	…	…	0.71
				Grand Average		**0.65 (0.06)**

M = Mission. Empty cells (…) indicate that mission was not attempted due to software problems. Standard deviation in parentheses

Table 11 Probability of high workload in the V-NIIRS condition for each mission

Participant	M1	M2	M3	M4	M5	Average
S1111	0.61	0.63	0.66	0.61	0.61	0.62
S2222	0.62	0.59	0.61	0.60	0.59	0.60
S0008	0.70	0.69	0.70	…	…	0.69
S0006	0.55	0.57	…	…	…	0.56
S0007	0.68	0.70	…	…	…	0.69
S0001	0.73	0.71	…	…	…	0.72
				Grand Average		**0.65 (0.06)**

M = Mission. Empty cells (…) indicate that mission was not attempted due to software problems. Standard deviation in parentheses

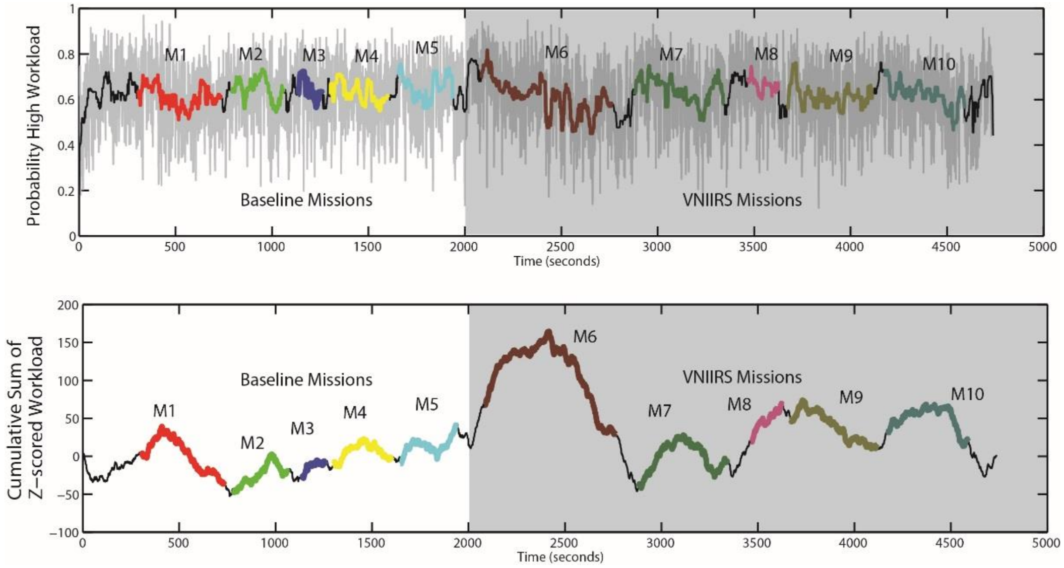

Fig. 14 Top: Continuous estimate of high workload probability over all missions (M) from S1111. Raw estimates are represented in light gray, and the black and colored segments are derived from a 29-s smoothing window. Bottom: The cumulative sum of the standardized workload estimates for all missions within the Baseline and V-NIIRS conditions.

3.3.2.2 Eye-Tracking

The eye-tracking data revealed that operators tended to make fewer blinks and more fixations on average in the V-NIIRS with respect to the Baseline condition (Fig. 15); however, this difference was not statistically significant. Figure 16 shows the gaze distribution from one subject during one of the missions presented in the V-NIIRS condition. The gaze data suggest this operator primarily searched for targets in the center of the video feed and continuously monitored or interacted with the timing parameters of the video.

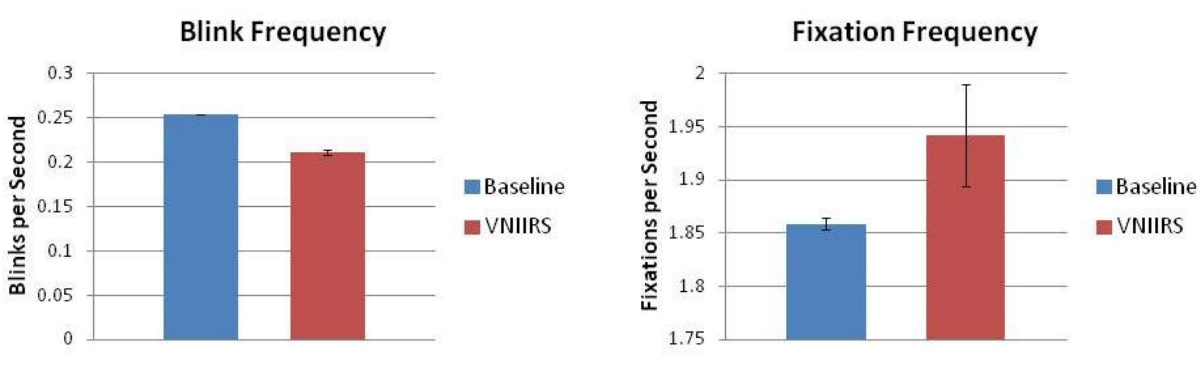

Fig. 15 Average blink and fixation frequency during target search across all analysts. Error bars equal standard error.

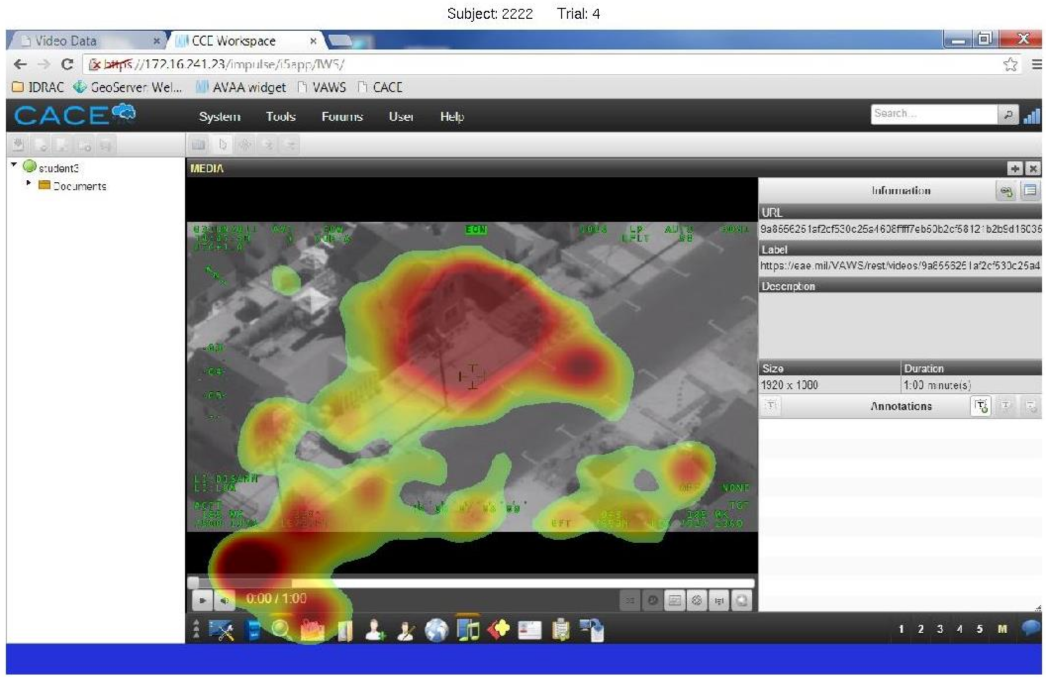

Fig. 16 Distribution of fixations from analyst S2222 during the fourth mission in the V-NIIRS condition. The video frame depicted is for illustrative purposes only.

3.3.2.3 Secondary Task Performance for EEG Participants: Auditory Probe Task

The operators made few errors when responding to the auditory targets presented in the secondary task. While there was little difference in the average accuracy to the targets between the Baseline and V-NIIRS conditions, the standard error was much larger in the V-NIIRS condition (Fig. 17). This was the result of one operator failing to respond to multiple auditory targets during one of the V-NIIRS missions. Reaction time to the targets was also similar between the Baseline and V-NIIRS conditions.

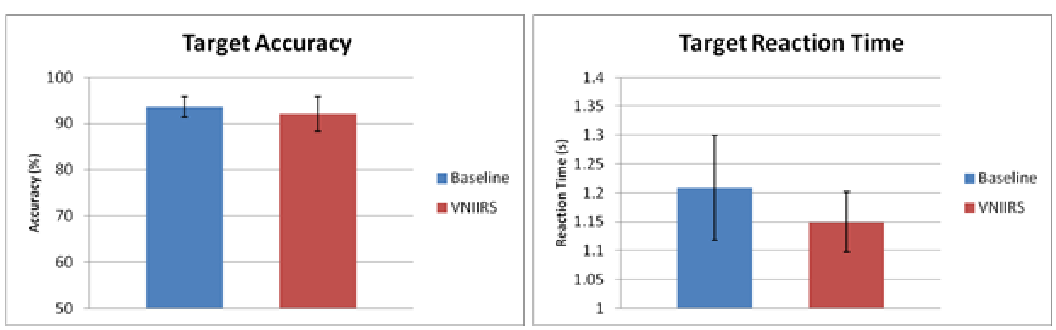

Fig. 17 Average accuracy and reaction time from all analysts to auditory targets presented in the secondary task. Error bars equal standard error.

3.3.3 Questionnaires

3.3.3.1 NASA TLX

The weighted workload ratings for the Baseline and the V-NIIRS condition are shown in heat maps in Tables 12 and 13, respectively. The warmer the color is, the higher the workload rating. The color patterns are very similar across the 2 tables. High workload ratings can be seen in Mental Demand (MD) and Performance (P). For the Performance scale, higher ratings are desirable, as they indicate that analysts were highly satisfied with their performance. Frustration (F) was generally low with one high rating of Frustration for participant 2. As expected, Physical Demand (PD) had consistently low workload ratings. The mean overall weighted workload rating was 8.42 (SD = 2.60) for the Baseline condition and 9.33 (SD = 3.83) for the V-NIIRS condition. The weighted workload for each category by condition is shown in Fig. 18.

Table 12 Heat map of workload ratings for Baseline condition

Participant	Instance	Condition	MD	PD	TD	P	E	F
1	2	Baseline	52	0	26	60	33	11
2	1	Baseline	24	0	12	36	10	15
3	1	Baseline	70	0	7	30	32	16
4	2	Baseline	48	0	27	40	6	6
5	2	Baseline	42	2	10	85	0	8
6	1	Baseline	6	0	4	90	4	6
7	1	Baseline	12	0	9	30	3	12

MD = Mental Demand; PD = Physical Demand; TD = Temporal Demand; P = Performance;
E = Effort; F = Frustration

Table 13 Heat map of workload ratings for V-NIIRS condition

Participant	Instance	Condition	MD	PD	TD	P	E	F
1	1	V-NIIRS	48	0	28	45	36	10
2	2	V-NIIRS	52	0	54	30	13	90
3	2	V-NIIRS	60	0	7	42	28	10
4	1	V-NIIRS	40	0	33	30	6	6
5	1	V-NIIRS	42	2	16	80	0	12
6	2	V-NIIRS	6	0	4	85	4	6
7	2	V-NIIRS	12	0	9	20	4	10

MD = Mental Demand; PD = Physical Demand; TD = Temporal Demand; P = Performance;
E = Effort; F = Frustration

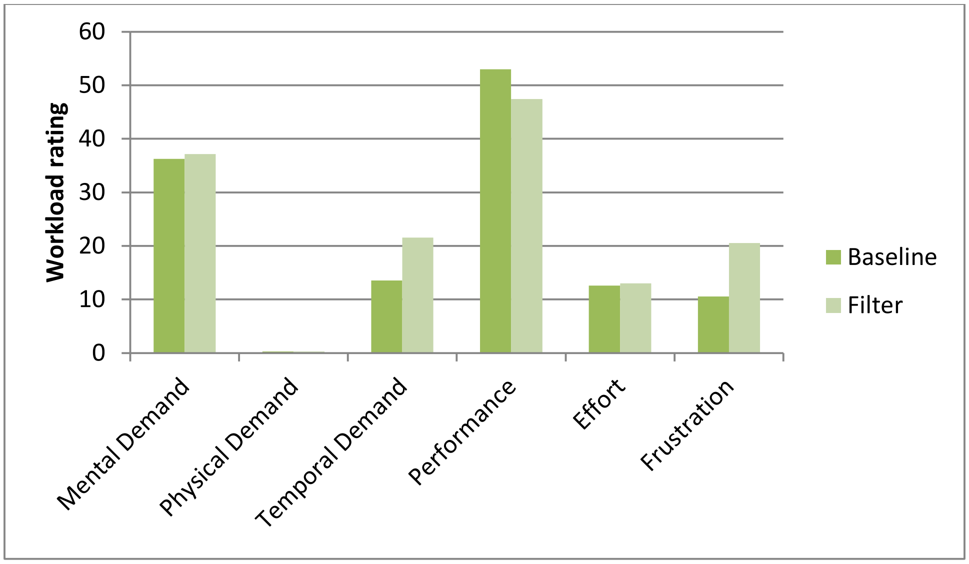

Fig. 18 Weighted NASA-TLX workload ratings by condition

3.3.3.2 Short Stress State Questionnaire (SSSQ)

Each subscale has 8 associated questions for a maximum possible score of 32. The mean Engagement score was 28.29 (SD = 1.98) with a range from 26 to 31. The mean Distress score was 4.57 (SD = 4.72) with a range from 0 to 13, and the mean Worry score was 8.00 (SD = 4.40) with a range from 3 to 15. Figure 19 shows the subscale scores by participant.

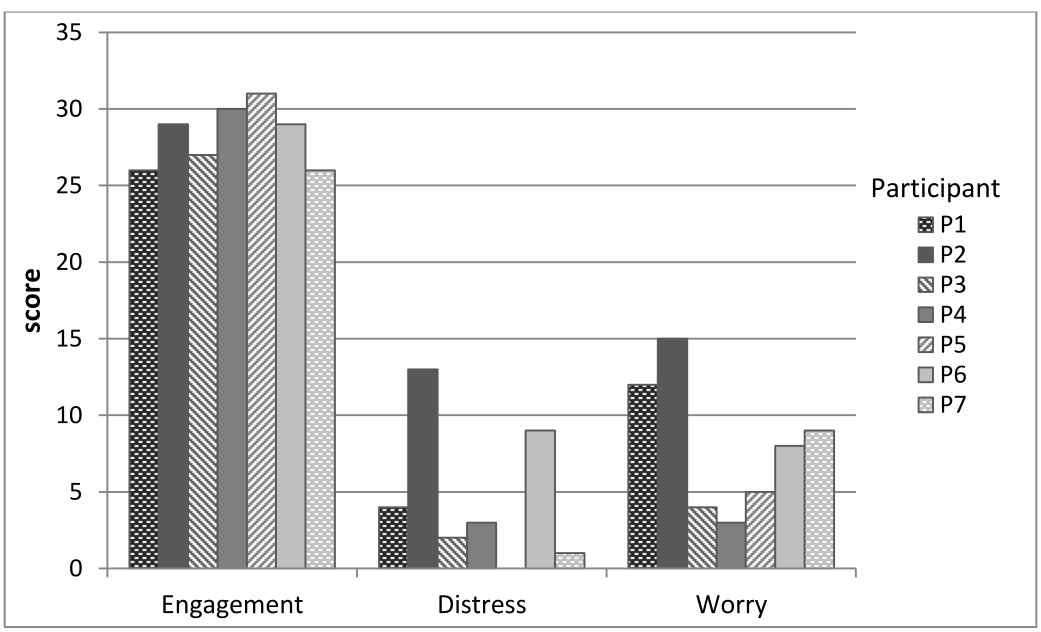

Fig. 19 Short Stress State Questionnaire (SSSQ) ratings for engagement, distress, and worry by participant

3.3.3.3 Usability

The average score over 43 items was 3.76 (SD = 0.87), comfortably in the positive direction. Overall, the percent of favorable ratings (e.g., a 4 or 5 rating) increased from 43% in the pilot study to 74%. The 17 questions in the Interface category had a mean of 3.67 (SD = 0.43). One analyst disagreed and another analyst strongly disagreed with the statement "Accidental keystrokes do not cause me to erase data or cancel a command." The following 5 statements had means greater than 4, with favorable ratings (a 4 or 5) from every analyst:

- The organization of the menus or information lists is logical.
- System information is presented in an understandable manner.
- Menu options are consistent in their wording, order, and location.
- On-screen instructions, prompts, and menu selections are easy to understand.
- It is relatively easy to move from one part of a task to another.

The 17 questions in the Functionality category had a mean of 3.52 (SD = 0.53). One analyst disagreed and another analyst strongly disagreed with the statement "When a keystroke (or mouse click) does not immediately produce the response I expect, the software gives me a message, symbol, or sign to acknowledge my input." Three analysts disagreed (2 of them strongly) with the statement "If AVAA rejects my input, it always gives me a useful feedback message (i.e., tells me why and what corrective action to take)." The following 5 statements had means greater than 4, with favorable ratings from every analyst:

- AVAA does not interfere with other programs I use.
- I can understand and act on the information provided.
- Data base queries are simple and easy.
- I can backtrack to the previous menu by using a single keystroke or mouse click.
- AVAA is easy to restart.

The 9 questions in the MANPRINT category had the highest mean at 4.16 (SD = 0.56). The following 6 statements had means greater than 4 with favorable ratings from every analyst:

- The number of personnel available in my unit/section is adequate to support full AVAA operations.

- I have the appropriate MOS to complete all assigned tasks.
- There are no physical limitations (color vision, hearing, etc.) that prevent me from completing tasks.
- The walk-through training gave me sufficient guidance so that I was able to complete my assigned task.
- Learning to use this software is easy.
- I feel confident in my ability to complete my assigned task using AVAA.

The 44th rating queried how long it would take before the analyst would be comfortable in the use of AVAA to complete job tasks. In general, the analysts felt that it would take less than a month to become comfortable with using AVAA in order to conduct their missions (a month was the lowest multiple choice item they could chose in that category).

3.3.4 Observations and User Comments

The debriefing comments are listed in Appendix C. Many of the user comments were consistent with the pilot study, including the ability to watch video faster than real time, differentiate videos in the video list, view overlays on the map, view the next video without returning to the video list, and view the history of annotation changes. Other comments were new. Users wanted to be able to view multiple videos simultaneously to save time and to compare the videos. Users indicated that it would be useful to zoom in to frames while the video was paused. They felt that automated tracking and automated annotations, in which the system identified potential targets and tracked the potential targets as they moved, would be beneficial.

4. Discussion and Conclusions

The sample size is not sufficient to conduct standard statistical analyses. The large standard deviations make it unlikely that statistical differences would be found. However, the expertise of the subject pool makes the data analysis and the insights they brought to the study worthwhile.

The V-NIIRS feature reduced the number of videos the analysts were required to search through and should be a worthwhile addition to AVAA depending on the actual military situation. For cloud applications, with multiple stored videos, it will probably be a necessity. The V-NIIRS-filtered FMVs were viewed to ensure that, at least for the experimental scenarios, the V-NIIRS filter did not screen out primary targets. Consistent with the pilot study, participants found 40% more primary targets with V-NIIRS. They also found 16% more total targets in the V-NIIRS compared to Baseline performance. This success with finding targets came at a cost of time. Participants were slower in finding primary targets with V-NIIRS, a difference of about 3.5 min. Thus, the V-NIIRS filter resulted in less work to be done, more targets found, but longer time to find the primary target. The longer time was due to the fact that in the baseline condition, the targets were found quickly but the operators did not find nearly as many targets as they did in the V-NIIRS condition. This suggests that in the baseline conditions, the analysts were only able to find the more obvious and thus the more rapidly acquired targets.

We again observed a large number of button clicks, primarily due to scrubbing the video to move forward within the video. Five out of 8 participants had over 7,000 clicks in 4 scenarios. Those who clicked more tended to make more annotations and find more primary targets, but it tended to take them longer to find the targets than those who clicked less. This shows that it is useful to move through the video quickly. The slower time may have been caused by the scrubbing process or it may have been caused by the fact that the participants took time to annotate other nonprimary targets. The implication is that it would be useful to have an automatic scrub feature that would jump through the video at intervals (thus reducing the need to click) or the ability to watch the video in faster than real time (to avoid missing any video frames). This would support the operator's workflow and greatly reduce the fatiguing number of mouse clicks.

One result that stood out was the fact that more experienced operators found fewer targets and took longer to find targets. They made fewer clicks and viewed fewer videos. Based on our observations, we believe this is because the more experienced operators were performing a more in-depth analysis of the FMVs. We hypothesize that they were considering and analyzing the full range of imagery on the FMV, not simply looking for one specific target. They were considering the terrain and the likely enemy actions and how they would create an intel product in context. In other words, they were treating the task more like they would a treat a true work assignment, not merely treating it as a simple experimental task. This deeper analysis caused them to take more time looking at the videos. We intend to better take advantage of the operators' experience in subsequent tasks. We will create tasks that involve not only annotating a target, but also creating an intelligence product based on the FMVs viewed. This will better tap into the analyst's ability to perform deeper analysis and will exercise AVAA on a more challenging cognitive task.

The analysts were all experienced combat Soldiers making their comments invaluable. Their comments and their survey evaluations indicate that AVAA should be a useful tool for the military intelligence community. We collected a number of useful comments about the usability of the tool and desired capabilities and features. In general, analysts found AVAA had operational utility and was easy to use. It is interesting that the number of usability statements with favorable ratings doubled between the pilot and the formal evaluation. This is likely due to the improvements within AVAA. The modifications to the experimental task (specifically, the addition of an operational context and a secondary target list) better exercised AVAA features, and there were fewer "not applicable" ratings on the usability questionnaire in the formal test.

Discrete subjective ratings from the NASA TLX were augmented with multiple continuous objective measures, including electrophysiology, eye-tracking, and behavioral performance. The measurement approach can be used in different environments and assess various cognitive states. The benefit of this approach is that it provides evaluators the ability to continuously track fluctuations in cognitive state during system interaction with higher temporal resolution than offered by traditional self-assessment approaches. This provides valuable information to evaluators in understanding how system implementations may impact cognitive state and, in turn, operator performance within the system.

Overall, the results indicated similar workload levels between the Baseline and V-NIIRS conditions. The NASA TLX data showed similar ratings between the 2 conditions for all demand factors. Behavioral performance was also similar between the 2 conditions as revealed by similar accuracy and reaction times to auditory targets. The eye-tracking data suggest a trend toward higher cognitive workload in the V-NIIRS condition as the blink frequency was lower and fixation frequency higher when compared to the Baseline condition; however, the auditory-evoked potentials exhibited N1 amplitudes of comparable magnitudes with slightly higher amplitudes in the Baseline condition.

There were instances when cognitive state derived from EEG correlated with subjective ratings and task performance. The notable rise in workload based on the classification of EEG from S1111 during Mission 1 of the V-NIIRS (Mission 6 overall) condition (Fig. 14) may have been reflective of frustration as this mission was ranked as producing the highest frustration (rated 52) in the V-NIIRS condition and was rated higher than the overall average rating of 29 on the NASA TLX. The N1 component of the auditory-evoked potentials revealed sensitivity to task engagement; however, it is difficult to ascertain the cause of the N1 amplitude difference (e.g., Fig. 13). For example, it may be that the particular mission in which the auditory probes were ignored produced significant workload demands on the operators such that they could not successfully perform both the visual search task and the auditory task together and thus automatically focused on the visual task. Alternatively, the operator may have intentionally ignored the auditory task and actively inhibited the auditory-evoked response. One caveat of using a secondary task is that it may negatively affect performance on the primary task; however, in the current test there was no apparent effect on the operator's performance in the visual search task while concurrently performing the secondary auditory task.

In conclusion, we have developed and implemented a multiaspect approach to estimate operator functional state during system evaluation. This approach is based on established scientific findings and provides evaluators a continuous objective means to estimate various cognitive states within a computer workstation environment. Further research must be done to validate this approach. It is critical that the validation process entails a large sample size, contains manipulations of various cognitive constructs that are easily manipulated and isolated, and provides high convergent validity between the measures.

5. Summary

Two data collection events at Ft. Huachuca were conducted: a pilot test and a data collection event. The sample size from either event is not sufficient to conduct standard statistical analyses. However, the descriptive statistics show trends of analysts being more successful but slower at finding targets in the V-NIIRS condition most likely due to far fewer (but more obvious targets) found in the baseline condition. For usability, the percent of favorable ratings (e.g., a 4 or 5 rating) increased from 43% in the pilot study to 74% in the June event. The expertise of the subject pool makes the data analysis and the insights they brought to the study worthwhile. The analysts were all experienced combat Soldiers, which made their comments invaluable. Their comments and their survey evaluations indicate that AVAA, even in its early configuration, should be a valuable tool for the military intelligence community. For cloud applications, with multiple stored videos, it will probably be a necessity.

We learned a number of lessons during the course of data collection. The small sample particularly for the EEG suggests we should not depend solely on the ICoE at Ft. Huachuca for participants. The ICoE analysts that participated were outstanding, cooperative, and thoroughly professional. However, a combination of exercises that are being held this summer at the EAE and the duty requirements of the analyst resulted in a smaller sample size than we would have wished. We are attempting to mitigate the problem by using analysts at the National Geospatial

Intelligence Agency outside of Washington, DC, and allowing a longer lag time to recruit participants for an early 2015 event in which we will compare new features (plug-ins) of AVAA to the baseline system.

6. References

Ahlstrom U, Friedman-Berg FJ. Using eye movement activity as a correlate of cognitive workload. International Journal of Industrial Ergonomics. 2006;36(7):623–636. doi:10.1016/j.ergon.2006.04.002.

Allison BZ, Polich J. Workload assessment of computer gaming using a single-stimulus event-related potential paradigm. Biological Psychology. 2008;77(3):277–283. doi:10.1016/j.biopsycho.2007.10.014

Backs RW, Walrath LC. Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. Applied Ergonomics.1992;23(4):243–254. doi:10.1016/0003-6870(92)90152-L.

Berka C, Levendowski DJ, Lumicao MN, Yau A, Davis G, Zivkovic VT, Craven PL. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviation, Space, and Environmental Medicine. 2007;78(5):B231–B244.

Dinges DF, Mallis MM, Maislin G, Powell I. Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management. Washington (DC): National Highway Traffic Safety Administration; 1998. Report No.: HS-808 762 [accessed 1998]. http://trid.trb.org /view.aspx?id=647942.

Dinges DF, Powell JW. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. Behavior Research Methods, Instruments, & Computers. 1985;17(6):652–655. doi:10.3758/BF03200977.

Federation of American Scientists. National Image Interpretability Rating Scales; 2014. [accessed 2014]. http://fas.org/irp/imint/niirs.htm.

Hart SG, Staveland LE. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock PA, Meshkati N, editors. Human Mental Workload. Amsterdam: North Holland Press; 1988.

Johnson RR, Popovic DP, Olmstead RE, Stikic M, Levendowski DJ, Berka C. Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. Biological Psychology. 2011;87(2), 241–250. doi:10.1016/j.biopsycho.2011.03.003.

Keller J. C4ISR: US military begins moving its information technology (IT) infrastructure to secure cloud computing. Military & Aerospace Electronics [accessed 2012 Jul 29]. www.militaryaerospace.com/articles /2012/07/dod -cloud-computing.html.

Kerick S, Ries AJ, Oie K, Jung T-P, Duann J, Chiou J-C, McDowell K. 2010 neuroscience director's strategic initiative. Adelphi (MD): Army Research Laboratory (US); 2011. Report No.: ARL-TR-5457.

Makeig S, Inlow M. Lapse in alertness: coherence of fluctuations in performance and EEG spectrum. Electroencephalography and Clinical Neurophysiology. 1993;86(1):23–35. doi:10.1016/0013-4694(93)90064-3.

MIL-STD-2525C. Common warfighting symbology. Arlington (VA): DISA Standards Management Branch; 2008 Nov 17.

Miller MW, Rietschel JC, McDonald CG, and Hatfield BD. A novel approach to the physiological measurement of mental workload. International Journal of Psychophysiology. 2011;80(1):75–78. doi:10.1016/j.ijpsycho.2011.02.003.

Peck EMM, Yuksel BF, Ottley A, Jacob RJK, Chang R. Using fNIRS brain sensing to evaluate information visualization interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York (NY): ACM; 2013. p. 473–482. doi:10.1145/2470654.2470723.

Stikic M, Johnson RR, Levendowski DJ, Popovic DP, Olmstead RE, Berka C. EEG-derived estimators of present and future cognitive performance. Frontiers in Human Neuroscience. 2011:5. doi:10.3389/fnhum.2011.00070.

Swett B. Advanced video activity analytics (AVAA) overview. AVAA Preliminary Design Review Presentations. 2013 Nov 6–7; Lorton, VA.

Van Orden KF, Jung T-P, Makeig S. Combined eye activity measures accurately estimate changes in sustained visual task performance. Biological Psychology. 2000;52(3):221–240. doi:10.1016/S0301-0511(99)00043-5.

Van Orden KFV, Limbert W, Makeig S, Jung T-P. Eye activity correlates of workload during a visuospatial memory task. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2001;43(1)111–121. doi:10.1518/001872001775992570.

Wang Z, Hope RM, Wang Z, Ji Q, Gray WD. Cross-subject workload classification with a hierarchical Bayes model. NeuroImage. 2012;59(1):64–69. doi:10.1016 /j.neuroimage.2011.07.094.

Wilson GF. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. The International Journal of Aviation Psychology. 2002;12(1):3–18. doi:10.1207/S15327108IJAP1201_2.

Wilson GF, Russell CA. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. Human Factors. 2007;49(6):1005–1018.

Appendix A. Forms and Questionnaires

This appendix appears in its original form, without editorial change.

Demographics

1. Age _________
2. Gender M / F
3. What is the highest level of education you received?
   (a) High school/GED (b) Some college (c) Bachelor's degree (d) Advanced degree
4. Rank ________
5. How many months or years you have served in the Armed Forces? ______________________________
6. What is your current MOS? ______________________
7. How many months or years have you had this MOS? __________________
8. What past MOS(s) have you held?
9. How many months or years did you hold each MOS?

______________________ → ______________________

______________________ → ______________________

______________________ → ______________________

10. If you have had training with imagery analysis other than your MOS training and duties, please describe it below. Include duration. _______ ____

    ____

11. How many months or years of experience do you have performing imagery analysis during operations (i.e. not training)? ____

12. If you have had experience with imagery analysis other than your MOS training and duties, please describe it below. Include duration. ______ ____

    ____

13. Have you participated in any previous AVAA experiments or familiarization? Y / N

    If yes, how many? ____

14. Do you wear eyeglasses or contacts regularly? Y / N

15. If yes, are you wearing them today? Y / N

16. How many hours of sleep do you normally get on a week night? ____________________________________

17. How many hours of sleep did you get last night? ____

AVAA Software Evaluation Date of Completion:______________________

The U.S. Army Research Laboratory is collecting data on your views about how well the Advanced Video Activity Analytics (AVAA) system meets user requirements. Mark the appropriate box for each question that supports your view of the system. Please explain all negative responses. If you have a comment or suggested improvement you can use the back of the page. Include the statement number and letter with your comment.

Comments should be as candid as possible since the ultimate goal of this evaluation is to provide the best system possible to the field.

A.	Rate the following statements related to the AVAA **interface**:	**Strongly Agree**	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**	**Not Applicabl e**
1.	The interface is free of unnecessary information.	☐	☐	☐	☐	☐	☐
2.	The organization of the menus or information lists is logical.	☐	☐	☐	☐	☐	☐
3.	I have no trouble finding and reading information on the interface.	☐	☐	☐	☐	☐	☐
4.	System information is presented in an understandable manner.	☐	☐	☐	☐	☐	☐
5.	It is easy for me to tell what data or files I am actually transmitting.	☐	☐	☐	☐	☐	☐
6.	Menu options are consistent in their wording, order, and location.	☐	☐	☐	☐	☐	☐
7.	On-screen instructions, prompts, and menu selections are easy to understand.	☐	☐	☐	☐	☐	☐
8.	Accidental keystrokes do not cause me to erase data or cancel a command.	☐	☐	☐	☐	☐	☐
9.	Audible signals (e.g., "beeps") help me avoid and correct mistakes.	☐	☐	☐	☐	☐	☐
10.	It is relatively easy to move from one part of a task to another.	☐	☐	☐	☐	☐	☐
11.	It is easy to change the way screen features such as icons are displayed.	☐	☐	☐	☐	☐	☐
12.	Data shown on the display screen are always in the format I need.	☐	☐	☐	☐	☐	☐
13.	It is easy to edit written documents, data entry fields, or graphics.	☐	☐	☐	☐	☐	☐
14.	If I make a data entry or typing error, it is easy for me to correct the error without having to retype the entry.	☐	☐	☐	☐	☐	☐
15.	The abbreviations, acronyms, and codes are easy to interpret	☐	☐	☐	☐	☐	☐
16.	It is always easy to tell what each icon represents.	☐	☐	☐	☐	☐	☐
17.	It is easy to acknowledge system alarms, signals, and messages.	☐	☐	☐	☐	☐	☐

B. Rate the following statements related to AVAA **functionality**:	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Not Applicabl e
1. AVAA does not interfere with other programs I use.	☐	☐	☐	☐	☐	☐
2. AVAA provides all the information I need to do my work.	☐	☐	☐	☐	☐	☐
3. I can understand and act on the information provided.	☐	☐	☐	☐	☐	☐
4. Data base queries are simple and easy.	☐	☐	☐	☐	☐	☐
5. The resulting operations of the numeric, function, and control keys are the same as for other tasks.	☐	☐	☐	☐	☐	☐
6. AVAA directs my attention to critical or abnormal data.	☐	☐	☐	☐	☐	☐
7. Importing data into the system is easy.	☐	☐	☐	☐	☐	☐
8. Exporting data out of the system is easy.	☐	☐	☐	☐	☐	☐
9. I can easily get a printed copy of the screen when I need it.	☐	☐	☐	☐	☐	☐
10. I rarely have to reenter data that I know is already available to AVAA in other files.	☐	☐	☐	☐	☐	☐
11. When a keystroke (or mouse click) does not immediately produce the response I expect, the software gives me a message, symbol, or sign to acknowledge my input.	☐	☐	☐	☐	☐	☐
12. Whenever I am about to enter a critical change or take some important, unrecoverable action, I must confirm the entry before accepting it.	☐	☐	☐	☐	☐	☐
13. If AVAA rejects my input, it always gives me a useful feedback message (i.e., tells me why and what corrective action to take).	☐	☐	☐	☐	☐	☐
14. I can backtrack to the previous menu by using a single keystroke or mouse click.	☐	☐	☐	☐	☐	☐
15. AVAA is easy to restart.	☐	☐	☐	☐	☐	☐
16. System log-on procedures are not unreasonably time consuming or complex.	☐	☐	☐	☐	☐	☐
17. System log-off procedures ask me if I want to save data before closing.	☐	☐	☐	☐	☐	☐

C.	Rate the following statements related to **manpower, personnel, training, and human factors engineering (MANPRINT)**:	**Strongly Agree**	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**	**Not Applicable**
1.	The number of personnel available in my unit/section is adequate to support full AVAA operations.	☐	☐	☐	☐	☐	☐
2.	I have the appropriate MOS to complete all assigned tasks.	☐	☐	☐	☐	☐	☐
3.	There are no physical limitations (color vision, hearing, etc.) that prevent me from completing tasks.	☐	☐	☐	☐	☐	☐
4.	The walk-through training gave me sufficient guidance so that I was able to complete my assigned task.	☐	☐	☐	☐	☐	☐
5.	Learning to use this software is easy.	☐	☐	☐	☐	☐	☐
6.	I feel confident in my ability to complete my assigned task using AVAA.	☐	☐	☐	☐	☐	☐
7.	Compared to my current method of exploiting imagery, AVAA does not affect my workload.	☐	☐	☐	☐	☐	☐
8.	Compared to my current method of exploiting imagery, AVAA decreases my workload.	☐	☐	☐	☐	☐	☐
9.	I have encountered no design or ergonomic issues with regard to system hardware.	☐	☐	☐	☐	☐	☐

9. How long do you think it took (or will take) before you consider yourself comfortable in the use of AVAA to complete your job tasks? (Please mark one)

Less than 1 month	2-3 months	4-6 months	7-12 months	More than 12 months
☐	☐	☐	☐	☐

10. What is the one thing you would do to improve the AVAA system?

____

____

11. Additional comments?

____

Raw Rating – complete after FIRST scenario

Please answer the following questions about your attitude **to the tasks you have just done**. Please place an "X" along each scale at the point that best indicates your experience with the display configuration.

Mental Demand: How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc)? Was the mission easy or demanding, simple or complex, exacting or forgiving?

Low High

Physical Demand: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the mission easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

Low High

Temporal Demand: How much time pressure did you feel due to the rate or pace at which the mission occurred? Was the pace slow and leisurely or rapid and frantic?

Low High

Performance: How successful do you think you were in accomplishing the goals of the mission? How satisfied were you with your performance in accomplishing these goals?

Low High

Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low High

Frustration: How discouraged, stressed, irritated, and annoyed versus gratified, relaxed, content, and complacent did you feel during your mission?

Low High

Part 1: Raw Rating – complete after <u>SECOND</u> scenario

Please answer the following questions about your attitude **to the tasks you have just done**. Please place an "X" along each scale at the point that best indicates your experience with the display configuration.

<u>Mental Demand</u>: How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc)? Was the mission easy or demanding, simple or complex, exacting or forgiving?

Low |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| High

<u>Physical Demand</u>: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the mission easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

Low |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| High

<u>Temporal Demand</u>: How much time pressure did you feel due to the rate or pace at which the mission occurred? Was the pace slow and leisurely or rapid and frantic?

Low |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| High

<u>Performance</u>: How successful do you think you were in accomplishing the goals of the mission? How satisfied were you with your performance in accomplishing these goals?

Low |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| High

<u>Effort</u>: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| High

<u>Frustration</u>: How discouraged, stressed, irritated, and annoyed versus gratified, relaxed, content, and complacent did you feel during your mission?

Low |_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_| High

Part 2: Weight – complete after the second scenario
This will be completed once after the second scenario. The weights will be used to calculate the total workload scores.

Directions: The evaluation you are about to perform is a technique that has been developed by NASA to assess the relative importance of six factors in determining how much workload you experienced. The procedure is simple: you are presented with a series of pairs of rating scale titles (for example, Effort vs. Performance) and asked to choose which of the items represents the more important contributor to workload for the specific tasks you performed in this experiment. Circle your choice.

Effort	or	Performance
Temporal Demand	or	Effort
Performance	or	Frustration
Physical Demand	or	Performance
Temporal Demand	or	Frustration
Physical Demand	or	Frustration
Physical Demand	or	Temporal Demand
Temporal Demand	or	Mental Demand
Frustration	or	Effort
Performance	or	Temporal Demand
Mental Demand	or	Physical Demand
Frustration	or	Mental Demand
Performance	or	Mental Demand
Mental Demand	or	Effort
Effort	or	Physical Demand

Scoring: An adjusted rating is achieved for each of the six scales by multiplying the weight by the raw score. An overall workload rating is achieved by summing the adjusted ratings and dividing by 15.

Stress: Short Stress State Questionnaire (SSSQ)

Please answer some questions about **the tasks you have just done**. Rate your agreement with the statements below by circling 4 for “extremely” agree, 3 for “very much” agree, 2 for “somewhat” agree, 1 for “a little bit” agree, and 0 for “no agreement at all”.

	Extremely	Very Much	Somewhat	A little bit	Not at all
1. I feel dissatisfied.	4	3	2	1	0
2. I feel alert.	4	3	2	1	0
3. I feel depressed.	4	3	2	1	0
4. I feel sad.	4	3	2	1	0
5. I feel active.	4	3	2	1	0
6. I feel impatient.	4	3	2	1	0
7. I feel annoyed.	4	3	2	1	0
8. I feel angry.	4	3	2	1	0
9. I feel irritated.	4	3	2	1	0
10. I feel grouchy.	4	3	2	1	0
11. I am committed to attaining my performance goals	4	3	2	1	0
12. I want to succeed on the task	4	3	2	1	0
13. I am motivated to do the task	4	3	2	1	0
14. I'm trying to figure myself out.	4	3	2	1	0
15. I'm reflecting about myself.	4	3	2	1	0
16. I'm daydreaming about myself.	4	3	2	1	0
17. I feel confident about my abilities.	4	3	2	1	0
18. I feel self-conscious.	4	3	2	1	0
19. I am worried about what other people think of me.	4	3	2	1	0
20. I feel concerned about the impression I am making.	4	3	2	1	0
21. I expect to perform proficiently on this task.	4	3	2	1	0
22. Generally, I feel in control of things.	4	3	2	1	0
23. I thought about how others have done on this task.	4	3	2	1	0
24. I thought about how I would feel if I were told how I performed.	4	3	2	1	0

Appendix B. Observations from the Pilot Study

This appendix appears in its original form, without editorial change.

SYSTEM FEEDBACK

Bugs

1. System occasionally froze on streaming video – appeared to happen with previously annotated video most often.
2. Clicking on a header in the video list to sort on sorts that page. It should sort all results.
3. Users should not be able to select an end date that is before the start date.
4. If search on date with 00:00:00 system only shows video for midnight. If you delete the time 00:00:00 the filter field still shows it.

Collected Capability Requests

5. Need fast forward/rewind and speed presets (double speed, x4, etc.).
6. There needs to be some way to differentiate the videos in the list. At a minimum date and time should be shown.
7. Need something on the video list (perhaps a different color or icon) that indicates a video has been reviewed/annotated (in session and in the past)
   a. Who looked at the video
   b. Has it been annotated?
   c. How much of the video has been played (similar to iTunes)
8. If an annotation is changed, notify those who previously used the annotation for a product
9. Ability to zoom in and out and pan from the mouse (scroll wheel), similar to Google Earth
10. Make it so that users can resize the window components (map, histogram, level of detail, tree view, etc.).
11. In real-time, mark an annotation without pausing video for another analyst to annotate or make a product
12. Ability to drag and drop MIL STD 2525 symbols onto video and have them geo-registered (need common symbols for annotations)
13. Ability to make video clips (extract a portion and make highlight video)
14. In the calendar widgets:
    d. Make the year and month drop-down options so users can either use the arrow buttons or select the month/year.

e. Once the begin date has been selected, default the end date to the same date (similar to the way airline sites work)

f. Do not allow the end date/time to be before the start day/time.

15. Add right-mouse menu to delete annotations.
16. Ability to automatically have the system go to the next video (or at least have a Next button so users do not have to go back to the list each time)
17. Ability to have shapes other than boxes for annotations (point, line, other shape annotations)
18. Ability to save frame as jpg or pdf
19. Ability to black out metadata or be able to pick what is shared (via a box or something)
20. Ability to switch from lat/long to MGRS
21. Ability to type any format of coordinates (lat/long or MGRS) quickly into search and have the map bring it up
22. Save a workspace – the map and FMVs currently working including the products created/under construction
23. Ability to customize the desktop/workspace area and have that saved with the user profile – which buttons, frames and other elements
24. Ability to save a video or set of videos to local system or server instead of working from the cloud for performance reasons.
25. Show the area the sensor is viewing FOV on map, not just the location of the sensor
26. Add quick search link or cookie crumbs to the video window that users can click to quickly get back to the search window (ex. Search -> Filter Search -> Search Results)
27. Ability for Date to be saved if move from “General” to “VAWS” filter search.
28. Ability to have map layers (like ArcGIS)
29. Ability to click on headers to sort.
30. Ability to highlight a group of video and have them play in sequence.
31. Ability to have search filter settings shown when playing the video.
32. Ability to see what platform shot the video.
33. Ability to search by platform (ex. Only show video shot by Hunter)

34. Ability to see timeline on annotation window.
35. Default map view should be of the world not any one particular area.
36. Ability to perform an advanced search on current set of results.

PROCESS FEEDBACK

37. The training before the actual exercises needs to be consistent across all groups.
38. During the exercises themselves the users should not give comments/feedback, they should concentrate on the tasks.
39. User feedback/comments should be collected at the end.
40. The "targets" need to be more detailed – several of the descriptions could be linked to items in the video's that did not match the target image.
41. We should think about adding an objective that is time limited, but allows users to find and annotate anything within a range that is potentially relevant. Measures would include number of videos reviews and number of annotations made.
42. Hide parts of CACE that are not relevant to AVAA and the experiment.
43. Operational context was missing. Potentially add something like "We just arrived in this area. Your goal is to survey a large area and find relevant activities, structures, and objects of interest using raw FMV that have not been surveyed before."
44. Investigate using CACE workflow feature for instructions.
45. Pre-test, time "playing" with the system should be a set time and the same for all users.
46. Need to clear annotations from free play time before starting experiment or have free play in a different geographic area or date/time than what is being used for the scenarios.
47. It would be nice to have a timer mechanism at each workstation – either a physical time the users can see or a program on the computer.

Appendix C. Observations from the June 2014 Study

This appendix appears in its original form, without editorial change.

SYSTEM FEEDBACK

Bugs

1. There was a “simple search bug” that sometimes occurred during a new search. The analyst entered time/date search criteria in the VAWS search but the simple search screen was automatically populated with other data, causing the system to crash or return the wrong videos.
2. Had one instance in which a big red bar showed up in the video. He had to go back and reload.

Collected Capability Requests

3. When on the map and trying to select a particular video, it takes multiple clicks to actually select the video. One click should highlight it, then the next should bring up the info.
4. Need the ability to watch the video in faster than real time (2x, 4x, 8x, 16x, etc.).
5. There needs to be some way to differentiate the videos in the list from each other.
6. The user should be able to tell which videos have already been viewed. Suggest using an icon that shows whether the video has been watched, partially watched, or not opened.
7. Increase the diversity and versatility of graphics that can be built during FMV exploitation. It would be nice to annotate using different shapes and colors than a blue box.
8. Ability to play multiple videos at one time, side by side. It would be a time saver, while one video is loading you can look at the other. It can also help in detecting changes.
9. Ability to click a button to play the next video without returning to the video list.
10. Provide error notes on why system has crashed.
11. Ability to zoom into frozen frames would be nice.
12. It is important to have track info when viewing video (map with video)
13. Annotation history should show who made changes and what the changes were.
14. Automatic tracking would be nice.
15. On the video list, it would be useful to see details such as the sensor platform, IR/EO mode, province, etc.

16. The option to have multiple selectable overlays is needed.
17. Automatic detection of objects or entities.
18. Would like to see geo rectified annotations.

PROCESS FEEDBACK

19. Having an overall operational context and list of secondary targets was successful. It was realistic, gave the analysts more to do, and provided another performance metric.

1 (PDF) DEFENSE TECHNICAL
INFORMATION CTR
DTIC OCA

2 (PDF) DIRECTOR
US ARMY RESEARCH LAB
RDRL CIO LL
IMAL HRA MAIL & RECORDS
MGMT

1 (PDF) GOVT PRINTG OFC
A MALHOTRA

1 (PDF) ARMY RSCH LABORATORY – HRED
RDRL HRM D
T DAVIS
BLDG 5400 RM C242
REDSTONE ARSENAL AL 35898-7290

1 (PDF) ARMY RSCH LABORATORY – HRED
RDRL HRS EA DR V J RICE
BLDG 4011 RM 217
1750 GREELEY RD
FORT SAM HOUSTON TX 78234-5002

1 (PDF) ARMY RSCH LABORATORY – HRED
RDRL HRM DG J RUBINSTEIN
BLDG 333
PICATINNY ARSENAL NJ 07806-5000

1 (PDF) ARMY RSCH LABORATORY – HRED
ARMC FIELD ELEMENT
RDRL HRM CH C BURNS
THIRD AVE BLDG 1467B RM 336
FORT KNOX KY 40121

1 (PDF) ARMY RSCH LABORATORY – HRED
AWC FIELD ELEMENT
RDRL HRM DJ D DURBIN
BLDG 4506 (DCD) RM 107
FORT RUCKER AL 36362-5000

1 (PDF) ARMY RSCH LABORATORY – HRED
RDRL HRM CK J REINHART
10125 KINGMAN RD BLDG 317
FORT BELVOIR VA 22060-5828

1 (PDF) ARMY RSCH LABORATORY – HRED
RDRL HRM AY M BARNES
2520 HEALY AVE
STE 1172 BLDG 51005
FORT HUACHUCA AZ 85613-7069

1 (PDF) ARMY RSCH LABORATORY – HRED
RDRL HRM AP D UNGVARSKY
POPE HALL BLDG 470
BCBL 806 HARRISON DR
FORT LEAVENWORTH KS 66027-2302

1 (PDF) ARMY RSCH LABORATORY – HRED
RDRL HRM AR J CHEN
12423 RESEARCH PKWY
ORLANDO FL 32826-3276

1 (PDF) ARMY RSCH LAB – HRED
HUMAN SYSTEMS
INTEGRATION ENGR
TACOM FIELD ELEMENT
RDRL HRM CU P MUNYA
6501 E 11 MILE RD
MS 284 BLDG 200A
WARREN MI 48397-5000

1 (PDF) ARMY RSCH LABORATORY – HRED
FIRES CTR OF EXCELLENCE
FIELD ELEMENT
RDRL HRM AF C HERNANDEZ
3040 NW AUSTIN RD RM 221
FORT SILL OK 73503-9043

1 (PDF) ARMY RSCH LABORATORY – HRED
RDRL HRM AV W CULBERTSON
91012 STATION AVE
FORT HOOD TX 76544-5073

1 (PDF) ARMY RSCH LABORATORY – HRED
RDRL HRM DE A MARES
1733 PLEASONTON RD BOX 3
FORT BLISS TX 79916-6816

8 (PDF) ARMY RSCH LABORATORY – HRED
SIMULATION & TRAINING
TECHNOLOGY CENTER
RDRL HRT COL G LAASE
RDRL HRT I MARTINEZ
RDRL HRT T R SOTTILARE
RDRL HRT B N FINKELSTEIN
RDRL HRT G A RODRIGUEZ
RDRL HRT I J HART
RDRL HRT M C METEVIER
RDRL HRT S B PETTIT
12423 RESEARCH PARKWAY
ORLANDO FL 32826

1 (PDF) ARMY RSCH LABORATORY – HRED
HQ USASOC
RDRL HRM CN R SPENCER
BLDG E2929 DESERT STORM DRIVE
FORT BRAGG NC 28310

1 (PDF)
ARMY G1
DAPE MR B KNAPP
300 ARMY PENTAGON RM 2C489
WASHINGTON DC 20310-0300

ABERDEEN PROVING GROUND

16 (PDF)
DIR USARL
RDRL HR
L ALLENDER
P FRANASZCZUK
K MCDOWELL
RDRL HRM
P SAVAGE-KNEPSHIELD
RDRL HRM AL
C PAULILLO
RDRL HRM AY
M BARNES
K SCHWEITZER
RDRL HRM B
J GRYNOVICKI
RDRL HRM C
L GARRETT
RDRL HRS
J LOCKETT
RDRL HRS B
M LAFIANDRA
RDRL HRS C
K MCDOWELL
A J RIES
J TOURYAN
RDRL HRS D
A SCHARINE
RDRL HRS E
D HEADLEY

INTENTIONALLY LEFT BLANK.